

Bridging the gulf

Ecologists and conservationists need to work more closely with economists and policy-makers if they are to make things happen on the ground.

Conservation biology is continually developing new tools and concepts that contribute to our understanding of ecosystems. In too many cases, however, that leaves scientists positioned only to track the loss of these systems. So far, researchers have been less effective at achieving the level of impact on policy decisions needed to implement actual conservation measures.

As long as this remains the case, it is hard to see how political pledges to conserve global biodiversity will be fulfilled. Under the 1992 Convention on Biological Diversity, for example, 188 nations are supposed to be taking steps to ensure that the rate of biodiversity loss slows down by 2010. But at the current rate of progress, it is hard to see how nations will reach even this modest goal.

The development of tools to monitor global biodiversity has helped to promote awareness of the scale of the environmental challenges facing the planet. But appropriate responses to these challenges are inevitably political and economic in nature. The considerable advances in monitoring and understanding made in conservation science cannot themselves generate such responses.

Translating the ramifications of environmental and conservation science into practical solutions requires much more work to close the gap between conservation biologists and the policy-makers and environmental managers who take action on the ground. One such effort is the RUPES programme run by the Nairobi-based World Agroforestry Centre, which is bringing together land managers, conservation groups, development agencies and researchers to design a system to reward mountain communities in Asia for the environmental services they provide by conserving local habitat.

If the drive for conservation comes only from scientists and a few allies in the environmental movement, ameliorative action won't get far. Economists and other policy-makers inside powerful government departments and development agencies are needed to design and develop plans to tackle the problem on a meaningful scale.

The most comprehensive survey yet of the economic and other benefits that natural ecosystems provide — the Millennium Ecosystem Assessment, published earlier this year — highlights the urgent need for closer dialogue between these different parties. The potential advances to be made from such discussion have never been more apparent. There is an increasing realization that economic arguments should be brought to bear in persuading policy-makers to protect environmental resources (see page 614). The United Nations and the World Bank are, at least in their public statements, stressing the potential of environmental conservation for improving quality of life in poor countries (see *Nature* 437, 180; 2005).

Putting these ideas into practice will require unprecedented collaboration between ecologists, economists, statisticians, businesses, land managers and policy-makers. As researchers continue to gather information about the kinds of benefits that ecosystems provide, it is critical that their findings are disseminated far beyond the scientific community.

This requires national institutions such as the US Department of the Interior, and international ones like the World Bank, to ensure that they have the necessary mechanisms and scientific expertise in place to absorb the information. Third parties, such as the H. John Heinz III Center for Science, Economics and the Environment in Washington DC, can also help to forge the necessary interactions.

A fuller dialogue will greatly benefit researchers, who can use it to establish exactly what kinds of information policy-makers and environmental managers need in order to translate science into effective action. Most of all, it will help the environment, by encouraging conservation policies that are soundly based on the facts. ■

"There is an increasing realization that economic arguments should be used to persuade policy-makers to protect environmental resources."

A missed opportunity?

Japan's prime minister has a valuable chance to reform his nation's tired scientific institutions.

This month's landslide re-election of Japan's Liberal Democrat government seems, on the face of it, to give Prime Minister Junichiro Koizumi a clear mandate to reform the country's institutions. One might reasonably expect that the universities and science agencies — whose performance today will help to determine Japan's technical and economic competitiveness tomorrow — would be near the top of the list. Unfortunately, there is scant indication that this rare opportunity will be grasped.

Japan's scientific and technical infrastructure is grounded in the two decades after the Second World War, when the country experienced rapid and remarkably successful industrialization. Its main elements are a proficient but profoundly conservative university system; a powerful civil service that briskly dispenses policy and priorities to the rest of the country; and a strong industrial research sector dominated by a handful of large corporations whose names have become synonymous with technical excellence.

This is a formidable combination that many other nations would envy — but, for the twenty-first century, it isn't enough. The system, however impressive in scale and scope, isn't flexible enough to take Japanese science to the next level, or to fuel the development of sectors, in biotechnology or computer software for example, that will fuel future economic growth. It is not set up to support research in



areas such as environmental and public health that match the non-economic aspirations of modern Japan. And it has demonstrably failed to impart Japan's government with the scientific know-how it needs if it is to assert badly needed regional leadership in Asia, on issues ranging from bird flu and global warming to the construction of large research facilities.

Unsurprisingly, none of this came up during the election campaign: Japanese politics rarely revolves around 'issues', in the Western sense. This time round, Koizumi's plans to reform the post office — the world's largest financial institution — were an exception to that rule. Politicians normally confine themselves to securing spending in the districts that they represent. Career civil servants, meanwhile, are systematically rotated between positions every two years and are sometimes more concerned with avoiding culpability than achieving results.

Scientific research has been popular with both politicians and bureaucrats primarily as a form of local spending, and it has been generously supported. Yet little thought has been given to its governance. This is one reason why Japan's scientific achievements are still falling some way short of its aspirations.

Too often, Japanese policy on important scientific issues is hammered out in back rooms. A public hearing is then held and a decision made. Outcomes are rarely clear-cut, and no one takes responsibility for implementing them. In the case of human embryonic stem-cell research, for example, researchers were told that they had the right to do it, but were so obstructed by red tape that little research has actually been done.

What could a genuinely reformist government do? It could start at

the grass-roots of science, in the universities, and make it a priority for them to open up both junior positions and tenured ones to young researchers, as well as to women and foreigners. It could introduce evaluation systems that encourage creativity instead of rewarding longevity. Some long-overdue changes at the universities, implemented last year, will have only a marginal impact on these issues.

The government should create an office, akin to the US Office of Research Integrity, to police scientific conduct. It should strengthen the Science Council of Japan, which advises the prime minister, and the Council for Science and Technology Policy, which influences the science budget, so the nation can develop a science policy worthy of its size and economic clout. It could fill some rank-and-file bureaucracy positions with scientists or former scientists, opening up a career path for struggling postdoctoral students. Currently the science ministry, the patent office and the main science funding agencies are all woefully short of staff with specialist knowledge.

Japan could then prepare itself to fill the leadership void in the Asia-Pacific region with regard to issues such as bird flu and global warming. It could then use scientific collaboration to improve relations with its neighbours, including China and South Korea.

There is little indication that Koizumi will do any of this. For as long as his government instead maintains its lukewarm embrace of science, Japan will continue to punch below its weight in terms of both scientific output and policy leadership in the region. ■

"Japan could use scientific collaboration to improve relations with its neighbours, including China and South Korea."

Do or die for design

A critical court case is addressing the teaching of 'intelligent design' in American schools.

This week, a federal court in Harrisburg, Pennsylvania, began hearing arguments about whether a school can promote intelligent design in the classroom (see page 607). A lawsuit brought by 11 parents of students in the Dover school district alleges that the local school board is violating the constitutional separation of church and state by requiring a statement promoting intelligent design to be read before teachers begin lessons on evolution.

Over the past few years, many scientists have worked hard to discredit intelligent design — but a favourable court verdict could damage the idea more than any amount of academic condemnation. For intelligent design was itself designed, in large part, to get around earlier court decisions that barred creationism from the classroom.

The first such ruling, by the Supreme Court in 1987, overturned a Louisiana law mandating that 'creation science', which sought to verify biblical creation through scientific enquiry, be taught alongside evolution. The second was a 1992 Arkansas finding that its very teaching violated the separation of church and state.

Intelligent design is a vaguer concept than creation science, and deliberately so. It posits only that an intelligent creator shaped the course of evolution. The general idea has been discussed by

theologians since Darwin's time, but it was only after these court rulings that it gained a significant following in the United States.

Unlike creation science, intelligent design is not affiliated with any specific religion. Rather than trying to prove its own explanation of the origin of species, it aims to punch holes in scientific doctrine. Its supporters, many of them fundamentalist Christians, have been hoping all along that the concept is sufficiently secular for the courts to permit its teaching in public schools.

If these hopes are realized, and the court rules in favour of the Dover school board, the movement is likely to spread quickly into many school districts. Political support for intelligent design, which has thus far been muted, would probably expand (see *Nature* **436**, 753; 2005).

But if the court rules in favour of the plaintiffs, this will seriously undermine efforts to get intelligent design into the classroom. What's more, Christian fundamentalists — some of whom are put off by intelligent design's ecumenical flavour — might then be inclined to abandon it for old-fashioned creationism.

Scientific organizations are well aware of this case's significance, and many have lent public support to the plaintiffs. A ruling in their favour will be welcomed not just by scientists and teachers but by American parents, whose children need to be protected from an injection of superstition into science teaching. ■

"If the court rules in favour of the Dover school board, 'intelligent design' is likely to spread quickly into many school districts."

RESEARCH HIGHLIGHTS

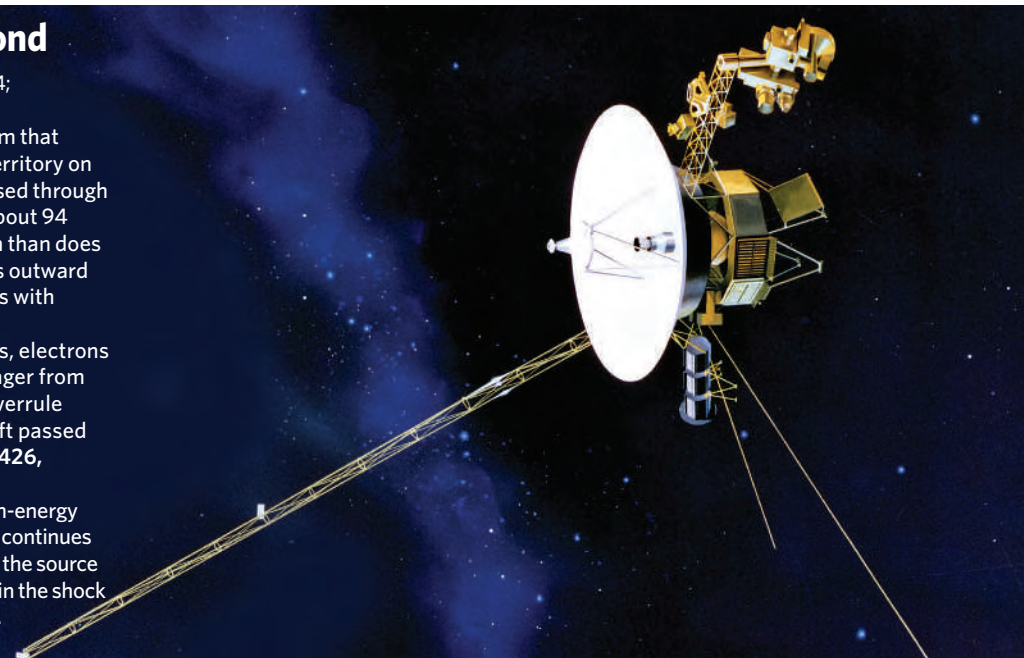
To infinity and beyond

Science **309**, 2017–2020; 2020–2024; 2025–2027; 2027–2029 (2005)

Data published last week confirm that Voyager 1 flew into uncharted territory on 16 December 2004 when it passed through the 'termination shock'. Lying about 94 times further away from the Sun than does the Earth, this is where the Sun's outward flow of charged particles merges with interstellar plasma.

The measurements of the ions, electrons and magnetic field around Voyager from four teams of US researchers overrule the previous report that the craft passed this milestone in 2002 (*Nature* **426**, 45–48; 2003).

Surprisingly, the number of high-energy cosmic rays detected by Voyager continues to increase. This could mean that the source of these rays lies beyond and not in the shock region as was previously thought.



NASA/JPL

MICROBIOLOGY

Kinky moves

Cell **122**, 941–945 (2005)

Video footage of swimming *Spiroplasma* has solved the long-standing puzzle of how these tiny helical bacteria move.

Some researchers believed that *Spiroplasma* travel by rotating their spiral-shaped bodies, like a corkscrew. They thought this because *Spiroplasma* lack the rotating, whip-like extensions called flagella that many other microbes use to move. But high-resolution video microscopy shows their motion to be more snake-like.

Joshua Shaevitz and his colleagues at the University of California, Berkeley, found that *Spiroplasma* move by unwinding their spiral shape from the front, then coiling it back up in the opposite direction. The resulting kinks that propagate along the body of this single-celled microorganism propel it forward.

CELL BIOLOGY

Inner charge

Proc. Natl Acad. Sci. USA **102**, 14058–14062 (2005)

A battery that stores electrical charge might build itself from a protein called Sprouty in the cells of mammals and other organisms, report Steven McKnight of the University of Texas Southwestern Medical Center and his co-workers.

Sprouty — so-called because fruitflies with mutant forms of it have excessively branched trachea — is thought to control development

by interrupting a cell-signalling pathway that regulates growth factors. While investigating the mechanism by which Sprouty works, McKnight's group noticed that sulphur atoms in the protein bind to iron, forming a complex that can hold and release electrons. The proteins clump into spherical particles 4–5 nanometres across, which might help to insulate the charge. The researchers suggest that such a particle could form the core of an unidentified enzyme.

NEUROBIOLOGY

Drugs to forget

Neuron **47**, 795–801; 873–884 (2005)

Two studies raise the prospect of a new treatment for drug addiction by showing, in rats, that it is possible to erase memories of cues associated with cocaine.

The treatments work by interfering with the pathways that reconsolidate a memory after its recall. Although such an approach has been shown to remove memories in other contexts, until now it was suspected that drug-linked memories might be too hard-wired.

In a study by Jonathan Lee of the University of Cambridge, UK, and his colleagues, rats were conditioned to associate a light signal with a cocaine reward. The animals' drug-craving response to light was eliminated by injecting the animals with DNA fragments that block the production of the protein Zif268. In the other study, researchers led by John Marshall of the University of California, Irvine, used drugs that block a biochemical pathway called ERK to erase the rats' preference for a chamber containing cocaine.

GENETICS

One too many

Science **309**, 2033–2037 (2005)

Geneticists have created the most accurate mouse model yet of Down's syndrome, a condition in humans caused by having an extra copy of chromosome 21. They did this by injecting mouse embryonic stem cells with copies of this human chromosome.

Previous mouse models had extra copies of parts of mouse chromosome 16, which bears many but not all of the same genes as the human 21. This meant that the full syndrome could not be studied.

The new mice show characteristics of

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

M. WEINZIERL/ALAMY

Down's syndrome, including behavioural changes and heart defects, reports the team led by Victor Tybulewicz of the National Institute for Medical Research, London, and Elizabeth Fisher of the Institute of Neurology, London.

PHYSICS

Feel the force

Nature Phys. doi:10.1038/nphys125 (2005)

A silicon chip that can juggle two blobs of ultracold gas provides a new tool for physicists exploring the quantum properties of Bose–Einstein condensates, and could form the basis of high-precision sensors.

The chip interferometer developed by Peter Krüger at the University of Heidelberg in Germany and his colleagues uses magnetic fields to split a condensate of rubidium atoms. The clouds of atoms are pulled up to 80 micrometres apart, such that there is interference between the quantum matter-waves of the two clouds.

Crucially, this separation does not affect the coherence of the condensates. This means that any changes in the way the two clouds interfere is a sensitive measure of external influences, such as a gravitational field, rather than an effect of the separation process.

MEDICINE

A good shot

J. Exp. Med. **202**, 817–828 (2005)

A dose of the drug chloroquine, delivered in conjunction with a vaccine, enhances the response of the immune system's CD8⁺ T cells. The finding, reported by Vincenzo Barnaba of the University of Rome 'La Sapienza', and co-workers, may represent a strategy to improve the effectiveness of vaccination.

Chloroquine reduces the acidity of the environment into which soluble viral antigens, the key components of many vaccines, enter when they are engulfed by a cell. This may slow the degradation of the antigens so that more are presented to the patrolling cells of the immune system, including CD8⁺ T cells, which then mediate an appropriate response.

CELL BIOLOGY

Bound by a ring

Cell **122**, 849–860 (2005)

Just before a cell divides, its chromosomes, which are organized as pairs of DNA molecules called chromatids, must be pulled apart so that each daughter cell can inherit one chromatid from each pair.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

A. T. SUMNER/SPL

Until this point, a protein complex called cohesin clamps the chromatid pairs (pictured) firmly together. The cohesin complex has recently been shown to be a large ring structure. Dmitri Ivanov and Kim Nasmyth of the Research Institute of Molecular Pathology in Vienna now show that the cohesin complexes seem to keep chromatid pairs together not by binding them physically, but by trapping them topologically inside their rings.

BIOCHEMISTRY

Stable mate

Nature Chem. Biol. doi:10.1038/nchembio734 (2005)

Although the nitric oxide produced by mammalian tissues is known to regulate cell function, the nitrite produced when it is oxidized was long viewed as biologically inert. Now a study by Martin Feelisch of the Boston University School of Medicine, Massachusetts, and his colleagues shows that nitrite can act as a signalling molecule and a regulator of gene expression.

Rats injected with nitrite showed marked changes in the activity of important enzymes such as cytochrome P₄₅₀. Nitrite can also set off a molecular cascade inside cells that ultimately affects blood-vessel dilation.

The authors note the similar action of nitrite and nitric oxide, and suggest that the overlap may offer an evolutionary advantage. As the more stable molecule, nitrite may act as a longer-lasting version of nitric oxide.

Corrections

Our Research Highlight 'Diamond geezers' (*Nature* **437**, 5; 2005) described a diamond material as "harder than the real thing". This is incorrect: the material is less compressible than diamond, as revealed through measurements of the bulk modulus. The reference for 'Keep your options open' (*Nature* **437**, 298; 2005) should have been: *Cell* **122**, 947–956 (2005). Apologies for the errors.

JOURNAL CLUB

Patrick Newell

Applied Physics Laboratory
Johns Hopkins University
Baltimore, Maryland

A physicist is drawn to wave research in his study of the aurora.

Little delights me more than work that unexpectedly unifies subjects previously thought disparate, particularly when it involves my specialty — the aurora.

Bright aurora, which form rings around the northern and southern magnetic poles, result from the impact of electrons on the upper atmosphere, some 120 km above the Earth's surface. The electrons which originate in the Solar wind or from the ionized layer of the atmosphere — the ionosphere — have somehow been accelerated to high energies.

For years, we focused on quasi-steady electric fields at heights of 1,500 to 10,000 km above the Earth's surface as the cause of the acceleration. More recently, evidence from satellites has suggested that some auroral electrons are accelerated by an entirely separate phenomenon: electromagnetic waves called Alfvén waves, which propagate through ionized gas.

Work in the *Journal of Geophysical Research* (C. C. Chaston *et al.* **110**, A02211; 2005) both solidifies this association and adds new wrinkles.

Chaston *et al.* show that regions where European Cluster satellites have measured a high flux of electromagnetic energy directed towards Earth — carried by an Alfvén wave — match up with areas where NASA's FAST satellite, in a lower orbit, has seen accelerated electrons. These electrons have just the type of energy spectra thought to correspond to wave-induced aurora.

Intriguingly, the paper also links the production of these Alfvén waves to surface waves on the magnetopause, which is the bubble that the Earth's magnetic field creates in the Solar wind. Although aurora remain my focus, I am now following wave research more closely.

NEWS



Marine environmental policy is being championed by a group of scientist advocates.

FLORIDA KEYS NATIONAL MARINE SANCTUARY

Scientists unite in bid to drive policy

SAN DIEGO

Tired of having their work ignored by politicians, scientists in the United States are taking matters into their own hands by using political organizations to advance scientific causes.

Political action committees, or PACs, have been around for years in US politics. They are typically used by powerful special-interest groups to collect donations while circumventing the controls on political contributions to specific candidates. But the handful of science-oriented PACs that have emerged in the past year or so represent a new trend to educate voters and politicians. They are formed by scientists and aim to influence voting or elected officials on specific topics, such as marine environmental policy, stem-cell research or conservation.

Ocean Champions, a California-based organization founded by marine biologist David Wilmot and environmental attorney Jack Sterne, is one example. "In the past, we would watch great science get ignored, manipulated or worse in the political process," says Wilmot, who has worked for several environmental groups. "We would have all our ducks lined up, but in the end we couldn't influence the political decisions. I was tired of losing. We are

now using science to create clout to drive good policy."

The PACs are typically targeting US congressional races — although some are already eyeing the 2008 presidential and state elections. And they say they will promote their causes, not any political party. For instance, of the 11 winners among the 14 Senate and House candidates backed by Ocean Champions last year, 6 were Democrats and 5 were Republicans.

Another science-related PAC, StemPAC of Washington DC, was created in July to push stem-cell research. StemPAC came out of 'kitchen table talk' by Democratic political consultants concerned about Tay-Sachs disease, for which stem-cell research might lead to therapies. The group jumped immediately into presidential politics by creating advertisements targeting Senator Bill Frist of Tennessee — the Republican majority leader of the Senate, and a physician, who had seemed reluctant to back stem-cell research.

The day the advertisements highlighting Frist's opinions were to begin running in New Hampshire, the site of the first presidential primary for the 2008 election, Frist publicly came out in favour of stem-cell

research. Political consultant Bud Jackson, a StemPAC founder, doesn't think the advertisements, which he says never ran, were the main reason Frist changed his mind, but says, "I think we contributed to hastening his decision." Politicians from all parties will be fair game, he adds. "If they are opposed to stem-cell research, we will hit them where it hurts."

StemPAC officials declined to discuss the organization's monetary goals. But at Ocean Champions, Wilmot says the organization aims to raise \$1 million during the forthcoming two-year federal-election cycle. Last week in Washington, the group held a 'coming-out' fund-raiser to boost its profile. Since forming in 2003, the group has raised about \$630,000, Wilmot adds.

Participation in such organizations may be a problem for scientists, many of whom are cautious about leaving the lab for the rough-and-tumble world of politics, and worried that their credibility or funding may be affected. "Scientists are afraid of advocacy," says ecologist David Blockstein, senior scientist for the National Council for Science and the Environment in Washington DC. "But this is changing."

Rex Dalton



GRAMMAR ANALYSIS REVEALS ANCIENT LANGUAGE TREE

It's not the words, it's how
you use them that counts.

www.nature.com/news

PUNCHSTOCK

Use of NIH funds placed under a spotlight

WASHINGTON DC

Congressman Joe Barton wants to know whether biomedical researchers funded by the US National Institutes of Health (NIH) are spending their grants haphazardly — overpaying research assistants, for example, or winning funds for phantom projects that they then use to do other research.

Barton, a Republican representative for Texas, heads the committee in the House of Representatives that oversees the NIH. After the committee's investigation of conflicts of interest inside the agency, revised ethics rules made their debut last month. Now the focus is on scientists outside the agency, at research hospitals and universities, who work with NIH funds.

After reading a 16 August article in the *Wall Street Journal* about a whistle-blower at

Cornell University's medical school in New York, and after receiving direct complaints, Barton's office sent two letters to Daniel Levinson, the inspector-general at the NIH's parent department, Health and Human Services.

One letter asks for a broad investigation into large grants to clinical-research centres, which can be worth many millions of dollars and cover many activities. The second asks for an investigation into whether NIH grant monies are being used to pay graduate research assistants unreasonably high salaries. This suspicion is based on complaints the committee received saying that some graduate assistants at the University of California, Davis, receive salaries and tuition waivers that amount to six times the salary of a postdoc.

Chris Harrington, director of communica-

tions in the University of California's federal-relations office, says he believes that the university complies with federal law.

And Norka Ruiz Bravo, deputy director for extramural research at the NIH, says she would be happy to cooperate with an investigation, but is not convinced there is a problem. "We're careful stewards of taxpayer funds. I would be surprised if there is widespread misuse of them," she said.

"You have to remember that these are grants, not contracts," she adds. "There is a certain amount of discretion left to the investigator on how to approach a scientific problem."

It remains to be seen whether Barton will agree. He has called for the investigation as his committee considers a draft of a sweeping reauthorization bill, which would affect the NIH's basic organization. ■

Emma Marris

"These are grants, not contracts. There's a certain amount of discretion."

Pioneering HIV treatment would use interference and gene therapy

Scientists have unveiled plans to test an HIV treatment based on a much-touted technique that hasn't yet been tried on people.

The treatment is based on a mechanism called RNA interference (RNAi), which can be used by cells to shut down invading viruses. Scientists and the biotechnology industry believe the interference pathway is a tremendously promising target for a variety of therapies. Two clinical trials of RNAi therapies have already begun, but the HIV proposal goes a step further, combining RNAi with gene therapy. It will be a closely watched test of whether the field can fulfil its potential.

Leaders of the trial described their plans to the US Recombinant DNA Advisory Committee (RAC) on 21 September. The committee gave generally favourable reviews, but recommended further safety tests before the study begins.

One of the trial's leaders is John Rossi, a molecular biologist at City of Hope's Beckman Research Institute in Duarte, California. Rossi says his team will perform

these extra tests before asking the Food and Drug Administration (FDA) for approval to begin the trial.

If the FDA says yes, Rossi and his team will test the therapy on five HIV patients who have a blood cancer called lymphoma. They will treat the patients' lymphoma with aggressive chemotherapy and a bone-marrow transplant — a normal procedure. But before the transplant, they will use gene therapy to add stretches

of DNA to stem cells in the bone marrow. It is hoped that molecules encoded by the added genes will trigger the cells' RNAi defences against HIV.

The trial is different from the RNAi trials already under way, because the molecules used in those studies remain in the body for only a short time. The City of Hope researchers will deliver DNA packaged into a gene-therapy vector that could persist in patients

for months or even years.

The RAC is cautious for that reason, and because the trial will set another precedent: it is the first in which researchers will use a lentivirus to deliver therapeutic genes to patients' stem cells. Lentiviruses are related to retroviruses, which were used in gene-therapy trials that caused cancer in three children with a rare immunodeficiency disease (see *Nature* 433, 561; 2005).

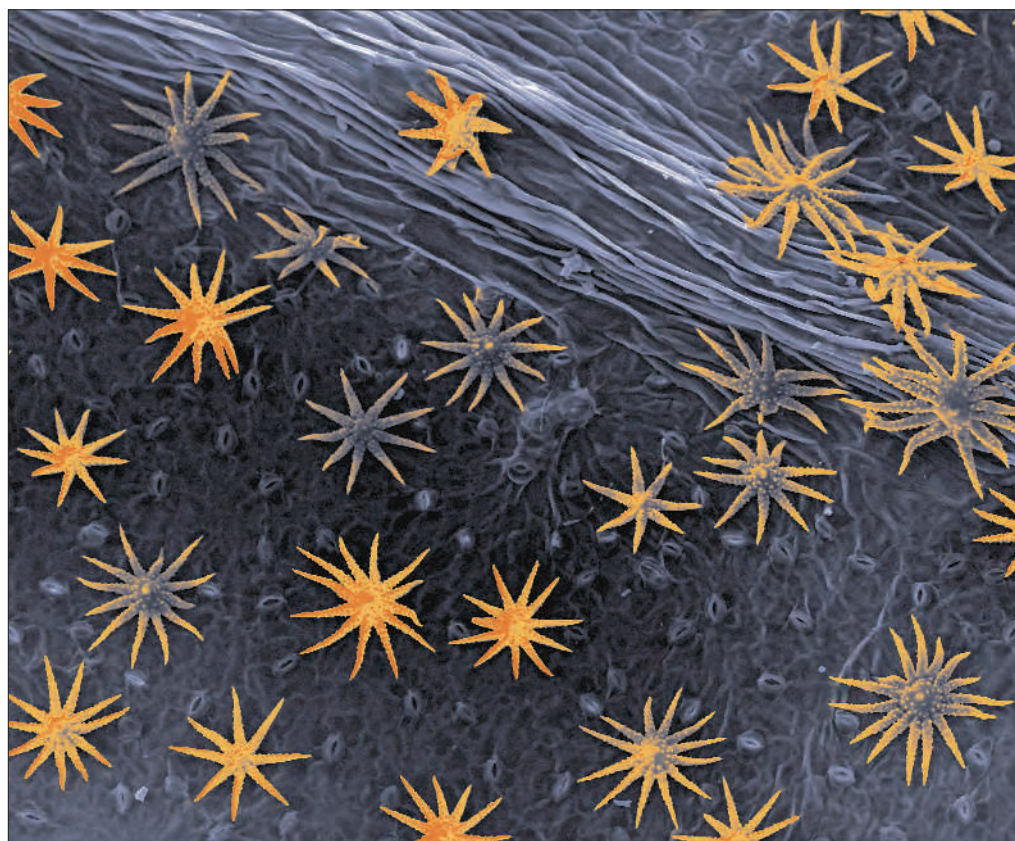
"We need to be careful, because now we have a study that's using a vector in the same family as the retrovirus, it's going into stem cells, and it's going into immunodeficient patients," says Diane Wara, who chairs the RAC.

The City of Hope team will monitor its patients to see whether the therapy causes cancerous mutations. The preliminary experiments are promising, Wara told *Nature*, emphasizing that she was speaking for herself, and not for the RAC. "John Rossi's work is beautifully done, and his data are very compelling." ■

Erika Check

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Interfering with HIV: a clinical trial of a combined treatment is under review.



SNAPSHOT Judges fall for a leaf's star quality

These are not orange starfish on the sea floor but tiny hairs on the underside of a plant's leaf. Known as trichomes, the hairs help the leaf to fend off the attentions of hungry insects and parasites. This image, taken using a scanning electron microscope, was highly commended in the Novartis and *The Daily Telegraph* Visions of Science competition, the results of which were due to be announced on 28 September.

The photographer is Stephanie Schüller of the Centre for Paediatric Gastroenterology at University College London, who captured the image while teaching a student how to use the microscope.

Europe tells Russia it faces HIV ruin

Health experts from Europe and the United States have called on the Russian government to strengthen its fight against the country's dramatically worsening HIV and AIDS problem. The epidemic there is set to explode, posing a serious threat to the former superpower's social and economic welfare, and even its stability, they told a parliamentary hearing at the Russian State Duma on 23 September.

"Russia needs to do a lot more than it has done in the past," says Chris McCafferty, chairwoman of a Council of Europe committee on social affairs that is launching a detailed study of HIV and AIDS in Europe. The committee is starting with Russia, and organized the Duma hearing as a first step. "It's bizarre," she says. "Sex is for sale on every corner here, but a sexual-health strategy just doesn't seem to exist."

AIDS arrived late in Russia, thanks to the country's relative isolation during the cold war, and the first case of HIV was not reported until 1987. But the epidemic is now believed to be growing faster in Russia and central Asia than anywhere else in the world.

According to official statistics, some 330,000

people have been infected so far, but little research has been done into the extent of the epidemic, and there is no credible reporting system for new cases. The number of unreported cases is thought to exceed 1 million.

As elsewhere, the disease first infected injecting drug users and sex workers, groups that often overlap. But it now seems to be moving into a second wave, where HIV is passed more by sexual transmission than shared needles.

The fear is that within five years, up to 10 million of Russia's 140-million population could be infected, says Alec Khachatryan, Russian programme director of Transatlantic Partners Against AIDS (TPAA), a New York, Moscow and Kiev-based charity that gave evidence at the hearing. Large-scale action is urgently needed to prevent the disease spreading from vulnerable groups to the general population, he says. "Neither AIDS prevention and treatment, nor research related to the disease, have been big priorities here."

In recent years the Russian government has

spent just US\$3 million to \$4 million a year fighting the disease. In comparison, Brazil, whose population is about the same size, spends almost \$200 million a year. Most Russian AIDS programmes are run by non-governmental and international organizations, such as the World Health Organization or the Bill & Melinda Gates Foundation.

Last week, the Russian Health Ministry proposed spending up to \$90 million extra each year. Among other things, this would provide people with AIDS with antiretroviral treatment, which can cost up to \$10,000 per person per year.

But Khachatryan feels that this would be far from enough. "What's really needed is a single AIDS strategy spanning research, prevention, treatment and human rights," he says. In an open letter to Russian president Vladimir Putin, the TPAA has asked for the creation of a presidential council on HIV and AIDS that would develop and review such a strategy.

Quirin Schiermeier

Political deadlock leaves scientists frustrated

MUNICH

While Germany's political parties struggle to set up a new government, scientists are left with little hope for more flexible regulations on stem-cell and biotechnology research.

The federal elections on 18 September led to a political impasse unique in Germany's post-war history. Neither Chancellor Gerhard Schröder's Social Democrat–Green coalition government nor the Christian Democrat–Liberal opposition led by Angela Merkel reached the necessary majority to elect a chancellor and form a government. As all parties have ruled out a coalition with the fifth force in parliament — the new Left Party — Germany's established political forces must hammer out another alliance.

At this stage, the most likely variant is a 'grand coalition' of Social Democrats (SPD), nicknamed 'red', and the 'black' Christian Democrats (CDU and CSU). But the rancorous power struggle between Schröder and Merkel over who would lead such an alliance is still threatening its creation.

A two-party government would be unlikely to produce any radical science policies, however. Under a grand coalition, research budgets would probably increase only modestly, as they did under the previous SPD–Green government — although the responsibilities for science and education could be simplified. Currently, these are split between central government and the 16 state governments.

Call for change

Such reform would give universities and research councils more financial freedom and is "urgently needed", according to Wilhelm Krull, secretary-general of the Volkswagen Foundation, Germany's largest private research funding agency. But for many researchers the priority is the country's strict rules on stem-cell and transgenic-plant research, which they believe are holding back science in Germany.

"There's clearly a need to change these overly restrictive laws," says Ferdinand Hucho, a biochemist at the Free University in Berlin. Hucho is also the main author of a recent report by the Berlin-Brandenburg Academy of Sciences, which highlights various German weaknesses in biotechnology and the life sciences.

The two major parties are deeply split over these issues, and few believe that things will get

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Gerhard Schröder (right) and Angela Merkel's parties are split over stem-cell research issues.

any better. In particular, the CDU candidate for the science ministry, Annette Schavan, a practising Catholic, is thought to be unlikely to touch existing restrictions in ethically charged areas of science.

The Liberal Free Democratic party (FDP), which won 10% of the vote, is the only party unambiguously in favour of stem-cell research — including therapeutic cloning — and plant biotechnology. Schröder has been courting the FDP heavily, but the 'yellow' party and its leader Guido Westerwelle have said they will not help a red–green–yellow 'traffic light' coalition into power.

Another possible, but unlikely, alliance is what has been dubbed a 'Jamaica' coalition of Christian Democrats, FDP and Greens. The implications for research are unclear, although most scientists admit that the Greens' seven-year involvement in power has been less of a problem for science than many expected (see *Nature* **436**, 1065; 2005).

Whatever government finally materializes, many scientists in Germany believe that support for specific directions in science is becoming a cross-party question, rather than the domain of single political groups. But they remain frustrated by the legal wrangling surrounding certain ethical issues. "The one thing I really wish," says Wieland Huttner, a director at the Max Planck Institute of Molecular Cell Biology and Genetics in Dresden, "is that science policies could be shaped by scientists, medics and engineers, and less by law experts."

Quirin Schiermeier

F. AUGSTEIN/AP

Science comes second as NASA makes lunar plans

WASHINGTON DC

NASA has set a goal of returning astronauts to the Moon in 2018 and begun defining a series of precursor missions for the next decade. But scientists and international partners are wondering how they fit into the plan.

NASA administrator Mike Griffin last week presented a broad outline of the programme, which he estimates will cost \$104 billion by 2018. On early missions, four astronauts will spend up to seven days on the Moon's surface, twice as long as the Apollo astronauts did. Eventually, crews will spend up to six months at a time living at a lunar outpost.

Long before people arrive, NASA will scout landing sites and do the technical groundwork with orbiting spacecraft and robot landers — starting with the Lunar Reconnaissance Orbiter (LRO) in 2008 and a landing mission targeted for sometime between 2009 and 2011.

The LRO, costing \$450 million to \$500 million, will be among the most capable planetary spacecraft ever built. Six onboard instruments will photograph the Moon and map its topography “at an engineering scale”, according to NASA chief scientist James Garvin. Five instruments will try to nail down the presence of water ice in shadowed craters — a critical resource for later human visitors.

Several nations are working on their own scientific missions to the Moon. Europe's SMART-1 probe, which has been in lunar orbit for nearly a year, recently had its mission extended to mid-2006. China, Japan and India plan to send orbiters before the LRO arrives.

But the rapid development of NASA's next two missions rules out international participation for the time being, Mark Borkowski, head of the Robotic Lunar Exploration Program at

NASA headquarters, told an international lunar-exploration conference last week in Toronto, Canada. “We are frankly disappointed,” he said. But in the long run, he promised, “we intend to go to the Moon with you”.

Planning for the first landing mission is under way. Teams from five NASA centres have proposed concepts, and the winner will be selected soon, says Borkowski. The lander is likely to be sent wherever the LRO detects ice — probably the south pole. It has yet to be decided whether it will be a \$450-million lander or a larger \$750-million version. If the latter, the launch could slip to 2011.

The current focus on hardware and site selection — engineering rather than science — worries some researchers. “My immediate reaction was: ‘So what are they going to do on the Moon? Where's the beef?’” says Wesley Huntress, a former NASA science chief now with the Carnegie Institution of Washington.

“You have to stick up for science all along,” adds Jeffrey Taylor, a University of Hawaii planetary scientist. He chairs the Lunar Exploration Analysis Group, set up to give NASA scientific advice on the Moon programme. It is identifying topics, ranging from biology to astronomy, that could form the basis of a research programme run by lunar astronauts or robots once a Moon base is in place.

But it is fine if science doesn't drive the programme, argues Lennard Fisk, a University of Michigan space scientist who chairs the National Academies' Sciences Space Studies Board. NASA has got into trouble by trying to justify projects such as the space station on the basis of science. “I'm relieved that we're not trying to force the science on the Moon,” he says. ■

Tony Reichhardt



Getting down to work: but will astronauts on NASA's planned missions be doing useful science?

ON THE RECORD

“We sent authors \$5 cheques. One altered it to \$6,005 and tried to cash it.”

William Gardner, of the University of Pittsburgh, on the fate of money sent to thank participants in a study of clinical-trial publication practices.

“If this makes the climate loonies in the States realize we've got a problem, some good will come out of a truly awful situation.”

John Lawton, chairman of the UK Royal Commission on Environmental Pollution, speaks out on the unusually fierce US hurricane season.

Source: *The Independent*

SCORECARD



Migrating birds

Spring and autumn will be darker in New York City. To reduce the number of birds hitting skyscrapers, lights above 40th floors must be turned off at midnight during migration.



Displaced dolphins

Hurricane Katrina threw eight trained dolphins from a Mississippi aquarium into the Gulf of Mexico. Marine biologists have now 'rescued' them back into captivity.



Mosquitoes

Scientists are developing a hormone therapy that aims to make mosquitoes urinate themselves to death.

OVERHYPED

Missing the win would hurt more

How bad is that pain? Medical emergencies aren't so urgent during a major sports event, as research has shown. Now scientists in baseball-crazy Boston have quantified the effect (J. S. Brownstein *et al.* *Ann. Emerg. Med.*, in the press). During league championships in 2004, when the local Red Sox looked set to lose, emergency-room visits were 15% higher than expected during such an event. By the time the Sox were winning the World Series (the first time since 1918), visits fell to 15% fewer than expected.

Hands on:
Andrew von Eschenbach
 will try to head two
 huge US agencies.

IMAGE
 UNAVAILABLE
 FOR COPYRIGHT
 REASONS

Cancer chief embraces top drugs job

WASHINGTON DC

Following the surprise resignation of its embattled commissioner, the US Food and Drug Administration (FDA) is now being led by Andrew von Eschenbach, director of the National Cancer Institute (NCI). He says that he will not give up his NCI job, despite the demands of both positions.

Lester Crawford was confirmed as FDA chief just two months ago, but he quit without explanation on 23 September. The agency has been buffeted by crises and has had a permanent commissioner for only 18 months of George W. Bush's administration.

Von Eschenbach will serve as interim commissioner until a permanent replacement can be found — or until President Bush nominates him for the job and the US Senate confirms him. In the meantime, he will remain at the helm of the NCI with its \$4.8-billion budget, an unprecedented dual assignment that has drawn criticism from some.

"It makes no sense at all. It's two huge jobs. Most officials have struggled to keep up with just one of those jobs," says a former senior FDA official who asked to remain anonymous. In addition, some are questioning whether von Eschenbach can ethically serve in both roles.

"As the head of the FDA, his job is to regulate

the companies that are bringing products before the agency to cure cancer. Isn't that a conflict of interest?" asks Merrill Goozner, head of the integrity of science project at the Washington-based Center for Science in the Public Interest.

But other experts disputed the notion that the two jobs conflict. At the NCI, "they absolutely do have a stake in what happens" in FDA drug-approval decisions, says Arthur Caplan, director of the Center for Bioethics at the University of Pennsylvania, Philadelphia. But that interest "is fuelled by public interest more than private gain", he says.

Von Eschenbach, a urology surgeon and cancer survivor, was a key leader at the University of Texas M. D. Anderson Cancer Center in Houston before becoming NCI director in 2002 (see *Nature Med.* **8**, 7 and 426; 2002). He has been closely associated with the US president's family, working with Bush and the first lady on the National Dialogue on Cancer — a public-private partnership that aims to develop a national cancer agenda. But at the NCI, he has displeased some cancer researchers by declaring an institute goal to eliminate death and suffering from cancer by 2015. Many in the field say that the goal is unrealistic and is setting up the public for disappointment.

Von Eschenbach could not be reached for comment by the time *Nature* went to press. In earlier statements he has said that he thinks promising drugs should be made available "as rapidly as possible". That approach is winning plaudits from industry: his appointment was praised by Amit Sachdev, who is the top health-policy official at the Biotechnology Industry Organization, based in Washington DC. The industry group Pharmaceutical Research and Manufacturers of America also hailed von Eschenbach's new role.

The reasons for Crawford's sudden departure were still mired in speculation early this week. Some newspapers reported that he had a financial interest that was not disclosed to senators when they were considering whether to confirm him as FDA chief. The confirmation battle took months and became caught up in a fight over whether emergency contraception should be made available without prescription.

Later, Crawford delayed a promised decision on the contraceptive, prompting a top official to resign (see *Nature* **437**, 179; 2005). Crawford was also the FDA's highest-ranking official when the painkiller Vioxx (rofecoxib) was withdrawn a year ago, raising questions about the agency's vigilance on drug safety. ■

Meredith Wadman

**THE STORM WATCHER**

Read our interview with the meteorologist in charge of predicting the course of US hurricanes.

www.nature.com/news

Into the eye of the storm

As Hurricane Rita headed for the US coast, **Mark Schroe** scored a rare trip into the gathering winds.

It's the up and down drafts that are the problem, I quickly learned while flying into Hurricane Rita. Crosswinds, even as fast as 250 kilometres per hour, aren't a big deal to a military P-3 airplane travelling 400 kilometres per hour. But a few of the convective drafts could have easily knocked me out of my seat if I hadn't been wearing a four-point safety belt. During one particularly turbulent moment, my free arm involuntarily flew as high as my head, giving some indication of what my internal organs were dealing with.

Throughout the hurricane season, the National Oceanic and Atmospheric Administration runs these stomach-churning reconnaissance flights from MacDill Air Force Base in Tampa, Florida. On 23 September, I was on a flight bound for Rita, just hours before the storm hit the Texas and Louisiana coast.

Such flights are the only way researchers can accurately gather key parameters on a hurri-



M. SCHROPE

Angry sky: the outer rainbands of Hurricane Rita seen from a plane approaching the storm.

cane, such as its central pressure and wind speeds. The pilots, experienced at flying into storms most others would consider deadly, use radar to thread a path through the outer bands and to target the central eye.

We made our first pass through the infamous eyewall, typically the most powerful part of a storm, just southwest of New Orleans. What followed was mostly a complete white-out, but after a couple of minutes of turbulence we broke out into the eye. The day before, I was

told, Rita had had a classic stadium-shaped eye, with clearly defined walls surrounding an open centre. Now, as the storm was weakening substantially from its former category-5 fury, the eye was poorly defined and filled with thin clouds. Quantifying this dramatic drop in strength was one of the flight's key discoveries. We were glad to give up the better show in exchange for a weaker landfall.

The eyewall passes caused the greatest flurry of activity on the plane. A technician across the aisle from me loaded data-gathering instruments called dropsondes into tubes and dropped them out of the plane, one by one.

Our departure from the eye after the first pass was much more jolting than the entry had been, as the western wall was more intense. It was on these turbulent passes that I came closest to 'earning my patch', as experienced fliers say. All passengers on the hurricane flights receive a commemorative patch, but veterans say you need a pounding to truly earn it.

Yet we knew our choppy flight was much smoother than what the residents of Texas and Louisiana, many of whose homes Rita would flood, would soon experience. While airborne, we heard a report that winds in Louisiana had already reached 60 kilometres per hour. Knowing firsthand what was to come, one crew member said respectfully: "Just wait." ■

School board in court over bid to teach intelligent design

HARRISBURG, PENNSYLVANIA

A US federal court this week began hearings on whether intelligent design deserves an airing in high schools as a viable scientific theory. The idea, which suggests that an intelligent creator shaped the course of evolution, is seen by many as an attempt to sneak creationism into the classroom.

On 26 September, scientists, legal experts, reporters and local families crowded into a ninth-floor courtroom in Harrisburg, Pennsylvania, to hear the opening arguments in the case. At issue is whether a school district in the nearby town of Dover has the right to require that intelligent design be mentioned in science classes.

The trial, observers say, is the most public airing to date on whether intelligent design should be taught in schools, and its outcome is likely to have ramifications for the teaching of science nationwide.

"This case is probably the most

important legal situation for creation and evolution in the past 18 years," Eugenie Scott, director of the National Center for Science Education in Oakland, California, said last week.

Last November, Dover's school board ordered that a short statement be read at the beginning of biology classes, which pointed to "gaps" in Darwin's theory of evolution and endorsed intelligent design as an alternative. Eleven parents filed suit against the district, claiming that the statement violated the required separation of church and state in lessons.

Eric Rothschild, a Philadelphia-area lawyer representing the parents, hammered home the point in opening arguments before Judge John Jones, who will adjudicate the suit. "The board changed the scientific curriculum to support a specific religious viewpoint," Rothschild told the judge. "And in

doing so they ignored the body of scientific knowledge."

But Pennsylvania standards require schools to teach students to think critically about scientific theories, and that's what the board's four-paragraph statement is designed to do, said Patrick Gillen of the Christian-oriented Thomas More Law Center in Ann Arbor, Michigan, who is representing the board. "This case is about free enquiry in education, not a religious agenda," Gillen said.

The plaintiffs' first witness was biologist Kenneth Miller of Brown University in Providence, Rhode Island, author of the biology textbook used in Dover classrooms. In more than three hours of testimony, Miller explained the scientific concepts of evolution and sought to show how they could explain things that intelligent-design advocates claimed were evidence of an intelligent creator.

"I believe that intelligent design is inherently religious," Miller told the judge. "And I think that the statement by the Dover board of education falsely undermines the scientific state of evolution theory."

During cross-examination, defence lawyers questioned Miller aggressively about the completeness of evolution theory. "The origin of DNA and RNA and their evolution is an unanswered question, is that correct?" asked Robert Muise, also of the Thomas More Law Center. Miller responded that some aspects of early DNA and RNA had been replicated, but that many questions remained unanswered. "I would rather say that Darwin was incomplete, not that Darwin was inadequate," he told the court.

The plaintiffs' testimony is likely to continue for at least a week before the defence takes over. A ruling is not expected until November. ■
Geoff Brumfiel

Cameroon's killer lakes defy gas extraction pipes

Lakes in Cameroon that can discharge large amounts of carbon dioxide still pose a health risk to local residents, despite pipes that are venting the gas from the waters, scientists say.

Some 1,800 people were asphyxiated when clouds of carbon dioxide belched from Lake Monoun in 1984 and Lake Nyos in 1986. The gas accumulates deep in the lakes, and erupts to the surface when enough has built up.

A degassing pipe was installed in each lake several years ago, but as gas is vented and pressure at the pipe inlets falls, the

removal process will slow down. George Kling of the University of Michigan and his colleagues predict that the pipes will remove about 30% of the remaining gas in Monoun, and 25% of the gas in Nyos, before they stop extracting gas faster than it is replenished.

Two pipes in Monoun and five pipes in Nyos would be needed to safely eliminate most of the remaining gas, the team says.

Database offers common ground for garden shrubs

Records of the living collections of 16 major US botanical gardens and arboreta are being collated in a database scheduled to go online late next year.

Called PlantCollections, the database will include information on 47,100 living plants from the institutions, and is being created using a \$666,000 federal grant awarded on 20 September.

"There is no comprehensive listing of all plants in cultivation — we envisage filling that void," says horticulturist Boyce Tankersley of the Chicago Botanic Garden in Glencoe, Illinois, who is coordinating the initiative. The new effort will not include row crops and trees, which can be searched through a government cataloguing system.

Once PlantCollections becomes fully operational, in three years, organizers hope to add reports of preserved specimens from the associated herbaria.

NASA axes role of chief scientist in shake-up

As NASA retools its workforce for a fresh round of Moon missions (see page 605), it is also changing the way it gets scientific advice. The agency's administrator, Mike Griffin, is scrapping the 12-year-old office of chief scientist. The current chief scientist, James Garvin, will move to NASA's Goddard Space Flight Center in Maryland.

Outside observers say that the decision shouldn't alarm the research community. "The position was always redundant," says Lennard Fisk, a space scientist at the University of Michigan in Ann Arbor. The chief scientist had no authority over the agency's \$5.5-billion science budget.

But the move comes at a time when NASA has a new and relatively untested head of space science — former astronaut Mary Cleave — and some worry that science will have a diminished voice within the agency. Compounding the problem is that NASA currently has no chartered science advisory

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Lake Nyos is being degassed, but not fast enough.

Analysis charts course to save ship's ancient timbers

PNAS

Efforts to conserve King Henry VIII's wrecked warship the *Mary Rose* received a boost this week from an in-depth analysis of the state of the ship's timbers.

Pride of the English fleet, the *Mary Rose* sank in 1545 near Portsmouth, UK. As a result, her timbers carry mineral encrustations formed over centuries under water, such as on the gun shield pictured right. A team led by Magnus Sandström of Stockholm University, Sweden, has studied the chemical changes that have taken place in wood from the hull, which was salvaged in 1982.

The team analysed sulphur compounds that had accumulated in the wood. Left untreated,



the sulphur could oxidize to form sulphuric acid and eat away at the timber. The findings should help conservation strategies to be revised, the team says (M. Sandström *et al. Proc. Natl Acad. Sci. USA* doi:10.1073/pnas.0504490102; 2005).

committee, although it put out a formal notice last week that this committee will be re-established.

India invests in research hub for biotechnology

The Indian government last week approved US\$7 million to establish a Regional Centre for Biotechnology Training and Education — a move intended to strengthen and focus biotech research and development across southeast Asia.

The centre, which has the backing of UNESCO, underscores India's desire to become a global player in the field (see *Nature* 436, 477–498; 2005). It will link biotech institutions in the region, aid technology transfer and develop doctoral programmes and training courses.

Officials say that once the centre is established, India will try to generate extra funds from United Nations agencies and other international organizations. Although a final decision has yet to be made, the centre is likely to be based in New Delhi, and could begin operating as early as 2006.

Australia reverses student visa cancellations

A flawed immigration notice has led to some 8,000 foreign students in Australia having their visas cancelled since 2001. As many as 300 students may have been wrongly detained and then deported, says the Department of Immigration and Multicultural and Indigenous Affairs (DIMIA).

In June, the Federal Magistrates Court ruled that the notice was flawed because it required students who had broken their visa conditions to report to a specific immigration office, rather than any DIMIA office or officer.

DIMIA announced on 16 September that it had reversed the visas automatically cancelled between May 2001 and 16 August 2005 for which the student did not report to the department within 28 days. The government is now trying to contact affected students to tell them that their visas have been reinstated.

Correction

The News Feature 'The nightmare before funding' (*Nature* 437, 308–311; 2005) gave the wrong point size for grant applications to the US National Institutes of Health. Applications must be written using text at a point size of 11 or more.

Mountain at the top

With one ageing telescope in space, and another mired in construction troubles on Earth, Matt Mountain has a tough job to do. **Jeff Kanipe** meets the new custodian of everyone's favourite space telescope.

Mattias Mountain seems cheerful as he sits at a desk littered with spreadsheets and organizational charts. This month he has become the director of the Space Telescope Science Institute in Baltimore, Maryland, at a time when the 25-year-old body is making an even bigger transition — from managing the popular workhorse of space astronomy, the Hubble Space Telescope, to its planned successor, the James Webb Space Telescope (JWST). Amid concerns that Hubble will be retired sooner than expected, and with JWST running behind schedule and over budget, the outlook for the institute seems far from rosy. Mountain admits that some of his friends have questioned his reasoning for taking the job but says he assures them: "I wouldn't have come here if I thought we were in our death throes."

Under a contract with NASA, the 400-person institute is responsible for research done with the \$1.5-billion Hubble telescope. When the space shuttle lofted Hubble into orbit in 1990, it launched a bold new era in observational astronomy — albeit after a false start. The incorrect curvature of Hubble's 2.4-metre primary mirror prevented light rays from converging at a single focus, blurring its vision. Only after a shuttle mission in 1993, during which astronauts installed corrective optics, did the bold new era truly begin.

And what a time it has been. In the ensuing years, Hubble has narrowed the age of the Universe to between 13 and 14 billion years, probed the violent hearts of galaxies, revealed dusty cocoons around newborn stars, helped to establish that cosmic expansion is accelerating and scoured the darkest, deepest reaches of space for primordial galaxies coalescing in a Universe less than a billion years old. Those were good times at the institute, which gained a reputation for smooth data management and slick public outreach. And beyond the pretty pictures, some 400 to 600 science papers are generated each year from Hubble-based data.

Hoping to keep Hubble working productively until 2010, the institute planned a fifth shuttle servicing mission to replace ageing



Matt Mountain hopes astronauts may once more extend Hubble's life (right), but technicians developing the James Webb Space Telescope will get no such second chances (bottom).

batteries and gyroscopes, and to add new instruments. But events outside the institute's control have made such hopes ever more remote. First, the disintegration of the Columbia orbiter during re-entry in 2003 led NASA to suspend shuttle missions. Then, this July, the shuttle Discovery narrowly escaped debris damage during the first shuttle launch since the Columbia disaster. In response, the new NASA administrator, Mike Griffin, has grounded the fleet for a second time.

Looking into a void

Although Griffin has said publicly that he is willing to reconsider a Hubble mission, time is running out. Even if shuttle flights resume next year, it is unlikely that a servicing mission would fly before late 2007. This may be too late to save Hubble should it lose battery power or if any more of its stabilizing gyroscopes should fail in the meantime. Just last month, NASA announced that engineers were shutting down one of the three remaining gyroscopes to extend the telescope's operating life until, possibly, mid-2008. Despite this, Mountain



NASA



NORTHROP GRUMMAN



remains optimistic about Hubble's future. "I think everybody thinks one shuttle servicing mission is a good idea, including the administrator," he says.

More critical to the future of the institute are the budgetary woes of the James Webb Space Telescope, projected to launch in 2011 at a cost of \$3.5 billion. Although not a direct successor to Hubble — JWST will observe mainly in the infrared — US astronomers picked the telescope as their top priority following a decadal review in 2000. The potential science to be done with the 6.5-metre infrared telescope is impressive: from observing the motions of young planetary systems around other stars to imaging the first ever galaxies to form in the Universe. But cost overruns are forcing astronomers to consider reducing the telescope's overall sensitivity and, perhaps, dropping other instruments entirely.

There is also a good chance that the launch date will slip to 2013. This could mean a long, dry period for the institute, whose staff are used to issuing a stream of data to astronomers worldwide. But Mountain is not worried about morale: "The spirits I detect around the corridors here are fairly upbeat. Now, I wouldn't say they're bubbling over, because NASA is putting budget constraints on this place, and that's a new experience for them."

As the former director of the Gemini Observatory, which operates two identical 8-metre telescopes in Hawaii and Chile, Mountain is no stranger to budgetary and instrumentation challenges. At Gemini, he learned that tight budget constraints can be another way of stimulating creativity. "I'm a great believer in the partnership between science, engineering and project management," he says, "To me it's a creative tension."

The big picture

Mountain will certainly need creativity to navigate the obstacles ahead. Craig Wheeler of the University of Texas at Austin, and president-elect of the American Astronomical Society, cautions: "I don't have any particular reason to think that he's not up to it, but I don't know whether he's faced quite this kind of challenge before. It's a big job." Many of Mountain's former colleagues are confident that he can pull it off, though. "One of Matt's real strengths is that he doesn't lose sight of the big picture," says Phil Puxley, in charge of Gemini's southern telescope in Chile. "He's delivered instruments and telescopes on a very tight schedule and under a lot of budget pressure. That's a very rare thing to do once, and he's now done it more than once."

The cost overruns, Mountain asserts, are not as bad as they seem. "If you ask a contractor what's the possible maximum cost," he says, "they always give you the worst-case scenario." When you add those contractors' estimates, he says, plus the cost of launch delays and NASA's estimates for unforeseen technical problems, the budget busts by a billion dollars. He argues the true figure for the overrun is closer to \$500 million — one that, fortuitously or not, matches savings that project scientists have recently put on the table.

Whatever the true amount, JWST's science working group, of which Mountain is a member, has worked hard since May to make savings while preserving as much of the telescope's performance as possible. One option, to reduce the overall size of the mirror array to 4 metres, was quickly rejected. Another option, to polish the mirrors once instead of twice, preserves the array's size but reduces its sensitivity at wavelengths shorter than 1.7 micrometres, the range that includes visible light. JWST was designed to operate at wavelengths between 0.6 to 28 micrometres, but it has always been primarily an infrared telescope, so the loss of the optical range doesn't concern Mountain. "Beyond 1.7 micrometres, JWST is supreme," he says. "My view is you don't need the optical."

Not everyone agrees. "We have certainly not run out of intriguing problems to be addressed in the optical range," says former Hubble project scientist Robert O'Dell of Vanderbilt University in Nashville, Tennessee. "The loss

of the optical band on the JWST will mean that there won't be a telescope that will do what we can right now. This loss will be serious."

Others warn of a potential observing 'gap' for certain observations, particularly if Hubble expires before JWST is operational. Astronomers say that further observations of type Ia supernovae, which they used to confirm that the Universe's expansion is accelerating, are needed to pinpoint exactly when acceleration began, and also to provide insights into the nature of dark energy, the force thought to be behind the acceleration. Losing the visible spectrum from 0.6 to 0.8 micrometres would not be a tragedy, says the institute's Adam Riess, whose team first announced the supernova result in 1998, but losing infrared from 0.8 to 1.2 would hurt without Hubble and "would leave us without any observatory to fill this niche". Mountain points out that JWST will have capabilities down to 0.6 micrometres, but without a second polish the performance will be degraded. He says astronomers will have to take what they can get at the shorter wavelengths.

While the JWST debate promises to dog Mountain's first year at the institute, he must

"The spirits I detect around the corridor here are fairly upbeat. Now, I wouldn't say they are bubbling over." — Matt Mountain

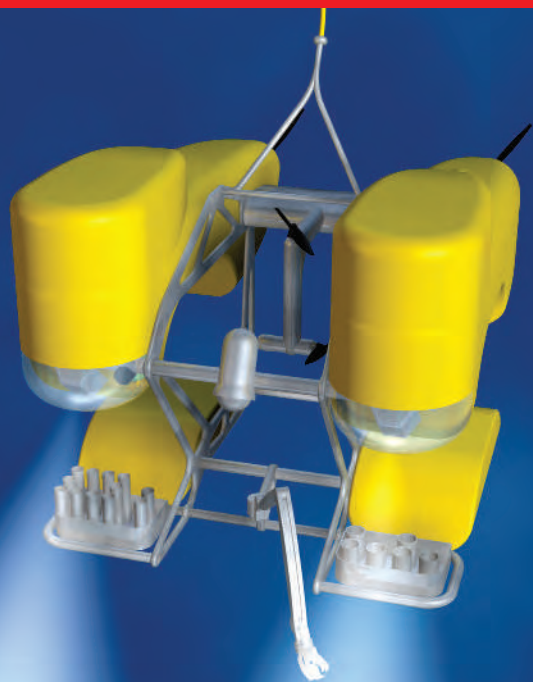
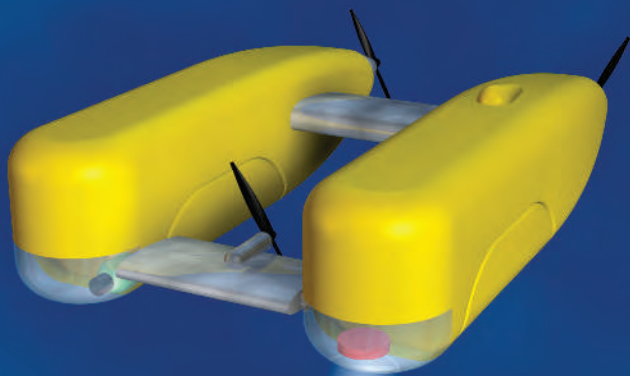
also contend with a political tug of war over the ageing Hubble. Mountain says he understands Griffin's reluctance to commit to a servicing mission

until there have been two safe shuttle flights. "It's the administrator's call," he says. "Both the current and previous NASA administrators have stressed that the decision is not a financial one but rather one of safety."

Griffin himself confirms that neither money nor safety is the primary concern. "The shuttle now operates under new constraints involving inspection requirements, use of spacewalk time and other factors which may limit its utility as a repair platform for Hubble," he explains. "It is too soon to be optimistic. If we fly the next mission in the spring of 2006, as we hope to do, and if all goes well, then we could be prepared to execute a Hubble mission in late 2007."

Although the current servicing mission is a big headache for the institute, JWST could be the source of many more, even once it is launched. Unlike Hubble, JWST will be stationed some 1.5 million kilometres from Earth, too far for rescue missions. This means the finished telescope must have no imperfections, nor parts that wear out fast. Mountain knows that the institute cannot rest on its laurels. "This is a very successful institution with a very motivated staff, and I think they feel that their past record justifies their continued existence," he says. "That's an understandable motivation, but it's not sufficient. We're going to have to earn our future."

Jeff Kanipe is a freelance writer based in Maryland.



Freestyle: the HROV submersible will be reugged from free-swimming mode (left) to tethered mode (right) aboard its mother ship.

Back to the bottom

Marine scientists are getting ready for their newest tool, a versatile robot submersible that can travel into the oceans' deepest abyss. **Robert Cooke** visits the Massachusetts lab where the future of deep-sea exploration is taking shape.

With a typhoon bearing down, the operators of the ship *Kairei* made what seemed the sensible decision: they hauled in their lines and planned to leave the area. Only these were no ordinary fishing lines, but a kilometres-long stretch of cable leading to the world's deepest-diving submersible. And when *Kairei's* crew winched up the last of the cable, something was missing — the vehicle on the end, whose line had apparently snapped.

The future of deep-sea exploration darkened a bit on that stormy day in 2003. The *Kairei's* submersible, *Kaiko*, was the star not only of the Japan Agency for Marine–Earth

Science and Technology but of the world's entire deep-diving fleet. In 1995, *Kaiko* had touched down on the bottom of the Challenger Deep in the Marianas Trench, 11,000 metres beneath the waves. It was only the second time a submersible had visited the legendary deep, and a first for a robotic craft.

But now, it looks as though *Kaiko* is to be replaced — and perhaps improved upon. At Woods Hole Oceanographic Institution (WHOI) in Massachusetts, home of the celebrated research submarine *Alvin*, marine engineers are preparing another vehicle to probe the deepest, darkest regions of the sea. Some components are still being designed, and even the overall shape has yet to be decided, but the engineers plan to have the vehicle in the water next summer and ready to dive to the Challenger Deep by 2007.

The craft, they say, will be much more than a replacement for the lost *Kaiko* or even for *Alvin*, which is due to be retired in the next four years or so. The new submersible will be a multi-talented, adaptable machine that can traverse the ocean in one of two configurations — self-guiding or controlled by surface operators through a whisker-thin fibre-optic cable. It's a combination of two successful marine exploration technologies — hence its unromantic designation of HROV, for Hybrid Remotely Operated Vehicle.

Like *Alvin*, but without a crew, the submersible will take pictures and video images, search with sonar signals, and grab rock and sediment samples off the sea floor. But on some missions, the HROV will also be sent off to go exploring on its own, in an autonomous mode for wide-area surveys.

Although the HROV will be the first submersible able to work either tethered or autonomously, the advantages of each mode are well proven. Autonomous underwater

vehicles have made a name for themselves in wide-field surveys and other tasks where constant monitoring is not required (see *Nature* **421**, 468–470; 2003). Freed from a tether, they can robotically 'fly' an underwater pattern to gather data on phenomena as different as hydrothermal vents and algal blooms. Remotely operated vehicles such as WHOI's *Jason II*, on the other hand, have the advantage that they can be steered to particular spots of interest, such as a newly formed hydrothermal vent. But the tethering cables produce drag, making these vehicles harder and slower to manoeuvre in the water.

Control and freedom

The \$5-million HROV is designed to have the best of both worlds. "We'll get a great deal of flexibility by being able to rearrange it," says Robert Detrick, vice-president for marine facilities and operations at WHOI.

In solo mode, the HROV will be programmed to follow instructions — to navigate through a search grid, for example — and then come back to its mother ship at the surface to deliver recorded data. If it locates something especially interesting, the HROV might be sent down again, tethered, to gather real-time data and respond minute-by-minute to commands from the surface. And it can do all that in a single day, with just one vehicle and one surface ship.

Better yet for marine scientists, the HROV will not be limited by depth. *Alvin* can safely reach 4,500 metres below the surface, which

"There is a very real risk. If one flotation ball implodes, the shock wave could take the whole vehicle out." — Andrew Bowen

JACK COOK, WOODS HOLE OCEANOGRAPHIC INSTITUTION

still leaves much of the ocean floor beyond its reach. The piloted *Mir* submersibles, owned by Russia, can reach 6,000 metres, as can the French submarine *Nautilus*. Since the loss of *Kaiko*, Japan has continued to explore the depths using its manned *Shinkai 6500* submersible, which can dive to 6,500 metres. The unmanned *Jason II* can also reach that depth.

Exploration of the deepest areas, however, must await the advent of the HROV. "The exciting thing is it should allow us to access areas that cannot be reached by existing manned vehicles," says Detrick. And there are discoveries to be made down there: when *Kaiko* visited the Challenger Deep, it brought back sediment containing primitive Foraminifera — single-celled organisms — that can withstand intense pressure.¹ The only other craft ever to visit the deep was the French-built *Trieste*, which carried Swiss scientist Jacques Piccard and the US Navy's John Walsh to the bottom in 1960.

But setting extreme-science records is not the HROV's goal, says Susan Humphris, a senior scientist at WHOI. Because of its multiple talents, the submersible will be useful at any depth. "One aspect I'm interested in is being able to access places we can't reach with current technology, especially in the Arctic," she says. Remotely operated vehicles generally aren't used in the Arctic for fear that ice might sever the tethering cables and the vehicle be lost. But with the HROV's dual capability, even if its cable is snapped it could still navigate itself back to its entry hole.

Because so little is known about the Arctic Ocean, any visit by the submersible is likely to yield important findings, says Humphris. She is particularly interested in scrutinizing the chemistry of Arctic seabed rocks: "These mantle rocks are the closest analogy we have today to what the whole planet's early materials were like," she says. Her group is already conducting preliminary experiments to test ideas about what geochemistry the HROV might find once it starts exploring beneath the ice.

Compared with other deep-diving submersibles, the HROV should be relatively easy to operate, in part because it won't need a dedicated ship and a large specialized crew to take it out to sea. Instead, all the necessary components will be in a van that can be lifted aboard almost any good-sized ship, with a small crew of dedicated technicians and engineers that will travel with it. The WHOI team aims to keep costs below \$10,000 per day, a third to a half as much as *Jason II* costs.

To keep costs low, much of the HROV technology will come off the shelf. The thin-walled ceramic spheres that will give the submersible buoyancy, for instance, are an offshoot of technology developed by the US beer company Coors, which worked on ceramics as a way to make better beer filters. And one key advance — the thin fibre-optic cable that links the HROV to the surface — was developed by the US Navy for torpedo guidance. The navy is



Buoyancy balls are being tested by researchers at Woods Hole.

collaborating on the project, as is robotics expert Louis Whitcomb of Johns Hopkins University in Baltimore, Maryland.

Batteries included

Unlike *Kaiko*, which received electrical power through its tethering cable, the HROV will be powered by onboard batteries. Because the cable serves only as a communications link and not as a power cord, it can be very slender and far more lightweight than *Kaiko's*, for example. To make sure it won't snap, the WHOI team has already tested it by running a live link between a surface ship and the sea floor, says Andrew Bowen, leader of the engineering team.

Plans for the vehicle are coming together at



The thin ceramic shells of the buoyancy balls are tested to well beyond deep-ocean pressures.

WHOI's Deep Submergence Laboratory at the institution's picturesque campus on Cape Cod. Not far from the busy ferry terminal crowded with tourists, engineers cluster in a large, garage-like chamber. Tools, machinery and parts are scattered across the concrete floor. This is where the HROV will be built.

But first, the engineers have to decide which new technologies to include.² "We're taking on the high-risk work early on," says Bowen. For instance, the team is experimenting with low-power light-emitting diodes to illuminate the dark depths. They are also testing different ways of keeping the vehicle buoyant, such as the extraordinarily strong ceramic spheres that use the Coors technology.

In the lab, Bowen displays some examples, white balls from 9 to 20 centimetres in diameter. Engineers are trying to decide which size would be best. The spheres are being tested to withstand pressures of 200 megapascals, even though they're expected to encounter only 120 megapascals at the bottom of the Challenger Deep. Such rigorous testing is necessary, Bowen explains, because the HROV can't afford to have a flotation ball break under pressure at depth. "There is a very real risk," he says. "If one implodes, the shock wave could take the whole vehicle out." Ceramic enclosures — either spheres or cylinders — are also being considered to house the HROV's instruments, guidance and electronic systems.

Whatever it ends up looking like, the HROV will fill a prominent empty space in oceanography's tool-kit. *Kaiko* may be lost, but soon the HROV will take over its job exploring the deep sea floor.

Robert Cooke is a freelance writer in Stow, Massachusetts.

1. Todo, Y. et al. *Science* **307**, 689 (2005).

2. Bowen, A.D. et al. *Mar. Technol. Soc. J.* **38**, 92-101 (2004).

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Dollars and sense

Approaches to conservation that seek to protect the most endangered species have had only mixed success. Is it time to move away from biodiversity 'hotspots' and stress the economic value of ecosystems? **Lucy Odling-Smee** investigates.

The Florida panther is living on the edge. Once, these majestic cats prowled throughout the southeastern United States. But today, fewer than 90 of the creatures cling to fragments of habitat in southern Florida. And not everyone agrees that efforts to save this subspecies make economic or scientific sense.

Male Florida panthers (*Puma concolor coryi*) stalk hunting grounds that average 550 square kilometres. Given the exorbitant cost of land in the Sunshine State, protecting sufficient habitat to support a population viable over the long term is a tall order. And although some argue that protecting the panther will rescue other threatened animals and plants along the way, this remains little more than an article of faith. Even the panther's evolutionary heritage has been called into question: genetic studies suggest that it is not as distinct from other subspecies of mountain lion as was once thought¹.

Attempts to save the Florida panther epitomize an approach to conservation that is increasingly coming under fire. A new, hard-headed breed of conservationists say we should not concentrate exclusively on saving

the rare and endangered or on protecting species diversity. Instead, they say, decisions need to be made within a rigorous economic framework. Some argue that the key to effective conservation is quantifying and promoting the economic 'services' that ecosystems provide for people — a mantra that has gained momentum with the completion this year of the most comprehensive survey yet of these benefits, the Millennium Ecosystem Assessment².

At the same time, conservationists are being urged to develop better tools to measure the effectiveness of their projects, and to share data on best practice. In other words, say critics, it's time for the organizations involved in conservation to admit that they are fallible, and to learn from past mistakes (see 'Taking quackery out of conservation', overleaf).

On the spot

In recent years, the field of conservation biology has been dominated by the goal of preserving biodiversity — a slippery concept, which can be defined in various ways. The most dramatic push came from an article³ published in 1988 by Norman Myers, then at Cornell University in Ithaca, New York. His

paper introduced the idea of biodiversity 'hotspots'. To earn hotspot status, Myers said, a region must contain 1,500 or more endemic plant species, which are found in that area but nowhere else, and it must have lost at least 70% of its original habitat. Myers identified ten areas of tropical forest as hotspots on the basis of these criteria.

It was a seductive idea: focusing scarce resources for conservation on hotspots offered maximum bang for buck. Conservation International, one of the leading organizations in the field, adopted the idea as its guiding principle in 1989. And subsequent analyses by Myers and others extended the concept from tropical forests to other habitat types and taxonomic groups⁴.

Conservation International, based in Washington DC, now recognizes 34 hotspots. These occupy just 2.3% of the Earth's land surface, yet are the sole home of half the world's vascular plant species and 42% of terrestrial vertebrates.

"Conservation International's maps have been an incredible political tool," says Ian Owens, a conservation biologist at Imperial College London. "They made rescuing biodiversity seem achievable."

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Counting costs: humans need vast forests (above), but these may be ignored if conservationists focus on the Florida panther (left) or hotspots in Peru (right).

But recently, the hotspot concept has come under fire. Analyses have revealed an alarming lack of overlap between hotspots identified using different criteria^{5,6} (see map, below). And some experts argue that focusing on biodiversity hotspots is fundamentally misguided. "It's like being a butterfly collector or having a zoo in which you protect a tiny sample of the Earth," says Peter Kareiva, a lead scientist for The Nature Conservancy, based in Arlington, Virginia. "Meanwhile, you could be ignoring ecosystems that are hugely important to humankind."

Hotspots are "questions waiting for answers," concludes Hugh Possingham, a mathematician and conservation biologist at the University of Queensland in Brisbane, Australia. He echoes Kareiva's call for emphasis on the importance of ecosystems to people, and wants conservation biologists to embrace the tools of decision theory. This theory is widely used in planning by engineers and financial advisers to work out how their funds

should best be allocated. Mapping more of the world's biodiversity hotspots is "like fixing the antenna on your car when the engine's broken", Possingham quips.

Possingham and his colleagues argue that spending money on those areas containing the most species at risk of extinction isn't necessarily the best strategy. Often, these are areas in which there is a small chance of success — because of overwhelming development pressure or official corruption, for example. In many cases the future of areas with fewer threatened species can be secured more easily and cheaply, he says.

In their current work, as yet unpublished, Possingham and his colleagues are using decision theory to lay economic factors over the maps of priority areas used by major conservation organizations. After plugging in the cost of action — which depends on factors such as land prices and human population density — their algorithms churn out an optimized strategy for allocating a limited pot of

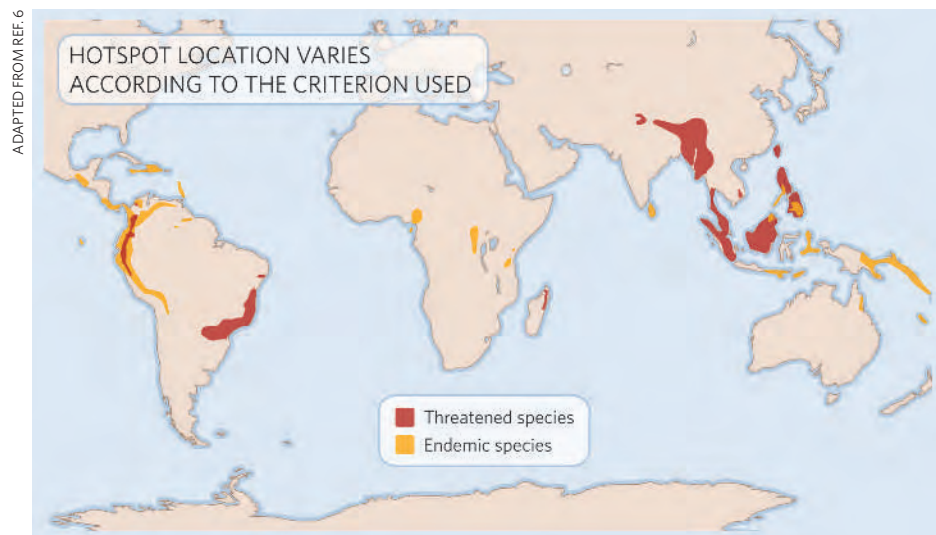
conservation funds. Kareiva sees Possingham's analyses as an early sign of a much-needed shift in thinking. "The whole conservation movement needs to deal more with people and with ecosystem services," he says. If it did so, he suggests, greater emphasis would be given to habitats such as the vast tracts of boreal forest that stretch from Russia to Canada. Nicknamed 'the world's lung', this habitat is an important carbon sink, providing a natural brake on the greenhouse effect, and it is arguably the planet's most important nitrogen-fixing ecosystem. Yet boreal forests are not a priority for several major international conservation groups.

Service not included

Natural ecosystems provide a wide variety of resources that have a social and economic value. These include services, such as clean water, stable soils and protection against natural catastrophes, and potential benefits, such as a storehouse of biodiversity from which drugs might be discovered. But studies to quantify these benefits, especially the financial costs and gains attached to protecting them, are only just beginning to gain momentum.

Preliminary results are eye-opening. Recent research indicates that the catastrophic loss of life seen in the Asian tsunami of 26 December 2004 could have been lessened had the clearance of Sri Lankan mangrove forests been prevented⁷. In Costa Rica, experiments have shown that maintaining a patch of forest, and so a supply of pollinators, near coffee plantations increases coffee yields by 20% — an economic gain that easily matches revenues obtained by converting the forest to farmland⁸.

At least now there is a solid base on which to build further analyses of the costs and benefits of protecting specific ecosystems: the Millennium Ecosystem Assessment. Requested by United Nations secretary-general Kofi Annan,



TAKING QUACKERY OUT OF CONSERVATION

Michael Wright knows he isn't being told the whole truth. The director of conservation and sustainable development for the Chicago-based MacArthur Foundation, Wright says he "falls out of his chair" if any of his grantees admit that their plans have misfired. "You get a proposal that says, 'Here are the things we want to do in the next three years', and then you get a report that says everything went according to plan," he says.

The true picture cannot be quite so rosy, Wright argues. But, in a field dominated by a few large organizations that rely on goodwill from foundations and the public to keep the money flowing, few conservationists are brave enough to admit to failure. "As much as fear of donors, it's institutional egos between organizations," says Wright.

In the past few years, efforts have been launched to make evaluation more rigorous and transparent. The three-year-old Conservation Measures Partnership, for instance, draws together big players in the field to create a common framework for deciding whether a project has succeeded. One goal is to harmonize terminology: for example, what Conservation International calls 'pressures' on a habitat or species, the WWF calls 'threats'.

But measuring the effectiveness of a particular project is only the start — the data must be disseminated to be useful. One attempt to do this is the website ConservationEvidence.com, run by William Sutherland of the University of East Anglia in Norwich, UK. The site accepts various accounts of how

interventions have gone, from journal articles to reports from wildlife managers. Several accounts of an issue are reviewed by an expert and encapsulated in an easy-to-read summary.

"I became increasingly uneasy about the fact that conservationists just make pronouncements about what is 'the right way'," explains Sutherland. He surveyed the people in eastern England who do the real work of conservation, such as park managers, and found that they get only 2.4% of their information from primary scientific literature⁹. His idea, he says, is to emulate the evidence-based medicine revolution launched in the 1970s, in which doctors began switching from tradition and intuition — and sometimes ineffective quackery — to remedies that

had been shown to work by scientific review.

The Centre for Evidence-Based Conservation at the University of Birmingham, UK, has similar goals. Since its launch in 2003, the centre has put out reviews on such topics as whether controlled burning of upland heaths helps to maintain floral diversity. "Conservation has stood still," complains Andrew Pullin, who heads the centre. "We're still making the same mistakes. Until we can get critical appraisal of our own actions, and make it available, we are not going to advance."

Nevertheless, both he and Sutherland are optimistic that their approach will eventually prevail. "I think we will cause a shift in the way conservation is done," Pullin predicts.

Emma Marris

it has been drawn up by more than 1,300 researchers from 95 nations over four years. It reviews the state of 24 different ecosystem services — from easily measured benefits, such as the provision of food, to elusive ones, which include the regulation of air quality and climate. Of these 24 services, 60% are being degraded, and fast².

Those involved in the assessment are disappointed with the response so far from the world's media and politicians. "If you went out and said we've looked at 24 indicators of economic well-being, and only four of them are improving, and of those four, one is about to crash, the world would panic," says Georgina Mace, a conservation biologist with the Zoological Society of London. The problem, she suggests, is that people aren't yet used to thinking about the environment as an economic resource.

Capital ideas

Nevertheless, the message is being picked up by influential figures within the conservation movement. Among the converts is Eric Dinerstein, chief scientist with the WWF, formerly the World Wide Fund for Nature, in Washington DC. "I don't think conservationists have sufficiently exploited the value of certain habitats that maintain services essential for human life and welfare," he says.

Eager to capitalize on this approach, the WWF is planning a scheme called 'hydrosheds'. This will use climate and hydrological models to identify the places where people get their water from. The goal is to produce a series of maps that can convince governments of the merits of conserving habitats that include economically important watersheds.

The practical difficulties of making such arguments work, however, are daunting. Andrew Laurie is chief technical adviser to a wetlands biodiversity project in China funded by the United Nations Development Programme and the Global Environment Facility. As well as providing diverse habitats for animal and plant species endemic to China, the

"The hotspot approach protects a tiny sample of the Earth. Meanwhile you could be ignoring ecosystems that are hugely important to humankind."

— Peter Kareiva

wetlands that Laurie is trying to protect act as water purifiers, floodwater and climate regulators, and suppliers of grass and reed building materials. Although the overall cost-benefit analysis gives a strong economic case for conservation, the equation is different for local farmers, who would lose the opportunity to convert wetlands for their own use. Devising specific financial mechanisms to reward these farmers will be key to success.

In Laurie's wetlands, there is at least a strong overlap between protecting biodiversity and promoting ecosystem services. Elsewhere, this isn't necessarily the case. Boreal forests, for instance, fare poorly on standard measures of biodiversity. And economic arguments relating to ecosystem services can, in some cases, usurp the goal of conserving wildlife. "You can cut down a mountain-top forest that has a lot of rare endemics, plant eucalyptus and probably get the same watershed benefit from the

introduced exotics as you would from native plants," Dinerstein admits.

Indeed, many conservation biologists are concerned that giving natural habitat a monetary value risks losing sight of the ethical and spiritual dimensions of conservation — driving forces in campaigns such as those to save the Florida panther.

"We mustn't rely only on an ecosystem-services approach because it misses out so much," says Laurie. "The argument that only by instilling respect for life are we going to get anywhere with conservation still carries a lot of weight."

But against the backdrop of environmental devastation now gripping the planet, and the scant resources devoted to conservation, there is a growing realization that economic arguments must become a key weapon in the movement's arsenal. "If conservation is to have any chance of being relevant in the next century," warns Kareiva, "it will only be because we have figured out how to protect ecosystem services at the same time as we protect biodiversity."

Lucy Odling-Smee is a subeditor for Nature.

1. Culver, M., Johnson, W. E., Pecon-Slattery, J. & O'Brien, S. J. *J. Hered.* **91**, 186–197 (2000).
2. *Ecosystems and Human Well-Being: Synthesis Report — Millennium Ecosystem Assessment* (Island Press, Washington DC, 2005).
3. Myers, N. *Environmentalist* **8**, 187–208 (1988).
4. Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. *Nature* **403**, 853–858 (2000).
5. Possingham, H. P. & Wilson, K. A. *Nature* **436**, 919–920 (2005).
6. Orme, C. D. L. *et al. Nature* **436**, 1016–1019 (2005).
7. Dahdouh-Guebas, F. *et al. Curr. Biol.* **15**, R443–R447 (2005).
8. Ricketts, T. H., Daily, G. C., Ehrlich, P. R. & Michener, C. D. *Proc. Natl Acad. Sci. USA* **101**, 12579–12582 (2004).
9. Sutherland, W. J., Pullin, A. S., Dolman, P. M. & Knight, T. M. *Trends Ecol. Evol.* **19**, 305–308 (2004).

BUSINESS

Appetite downer awaits approval

A pill that works by putting the hunger induced by cannabis into reverse could jump-start a languishing market for obesity drugs, reports **Meredith Wadman**.

A French drug company could be within months of winning regulatory approval for rimonabant, a drug widely tipped as the first 'blockbuster' weight-loss pill.

After decades of false starts, the prospect will raise a sceptical eyebrow or two. Previous weight-loss therapies have left in their wake a trail of disappointed dieters and dangerous or unpleasant side effects — not to mention an obesity epidemic that is veering out of control.

But Paris-based Sanofi-Aventis said in June that the US Food and Drug Administration (FDA) is considering its new-drug application for rimonabant, which would go on sale in the United States as Acomplia. A decision could come as soon as next spring. The company, which was formed in a merger last year and is the world's third-largest drugmaker, has also applied for approval in Europe and elsewhere.

The drug works by blocking one of the receptors where the active ingredient in cannabis docks. Dubbed the CB1 receptor, it is widely dispersed in the brain and other organs. The receptor helps to regulate energy balance, fat and sugar metabolism, and appetite, and is stimulated by the body's own cannabis-like neurotransmitters. In what smokers of the weed might recognize as a 'reverse munchie' effect, the blocking of the receptor makes people feel satiated, so they eat less.

Impressive results

In one completed clinical trial, investigators reported in April that 363 obese Europeans who took 20 milligrams of the drug daily for a year lost an average of 8.6 kilograms (*Lancet* 365, 1389–1397; 2005). Their waistlines shrank by an average of more than 8 centimetres. And this wasn't just a cosmetic improvement: excess abdominal fat is a significant risk factor for heart disease.

The subjects also showed healthy changes in blood levels of important heart-disease-related substances: levels of HDL, the so-called 'good' cholesterol, rose significantly, and those of triglycerides — fat transport and storage molecules — fell. The most common side effects included nausea (12.9%), dizziness (8.7%) and diarrhoea (7.2%). But investigators said that these were "mild to moderate and considered to be transient, based on the occurrence mainly during the first months of the study".

Over the past 18 months, researchers have



reported other findings from a total of seven phase III clinical trials involving over 13,000 subjects. These indicate that the drug is effective not just for weight loss but also in controlling diabetes and even in quitting smoking.

These results make Sanofi far and away the leader of the pack of drug companies chasing the potentially vast obesity market. Scientists in company laboratories at Montpellier in France began their hunt for the drug soon after the brain receptor for cannabis was identified in 1990 (see *Nature* 346, 561–564; 1990). Their reasoning was that if marijuana creates the munchies, a compound that blocks its effects could decrease appetite. That early insight has

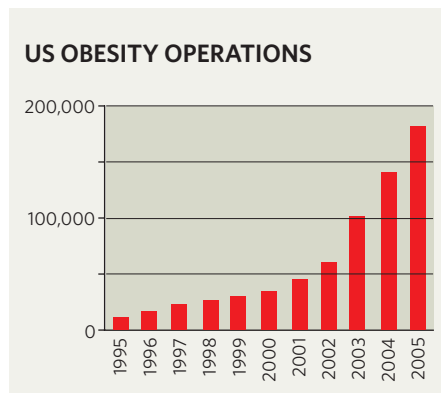
left more recent entrants to the field, such as Merck, Pfizer and Bristol-Myers Squibb, scrambling to recover lost ground.

Not for dieters

But the French company has been at pains to portray rimonabant as anything but a pill for cosmetic dieters. "There have been a whole host of positive impacts we've seen with this product in overweight, obese and diabetic patients," says company spokeswoman Julissa Viana. "It's not for someone who just wants to lose five or ten pounds." That being said, physicians in the United States can prescribe an approved drug to anyone for any purpose.

Sanofi's market timing could hardly be better. Last week, the World Health Organization declared that more than 1 billion people worldwide are overweight, a number expected to grow to 1.5 billion by 2015. Yet in the United States, only three key weight-loss drugs are currently approved; their annual sales totalled \$224 million last year, according to IMS Health, a Pennsylvania-based pharmaceutical information and tracking company.

"As a business, this is just terrible," says Jose Caro, an endocrinologist who is in charge of obesity drug research at Eli Lilly, based in Indianapolis. He says his company has five anti-obesity compounds of its own in early



J. RAEDLE/GETTY

SOURCE: AMERICAN SOCIETY FOR BARIATRIC SURGERY

clinical development. According to Decision Resources, a market-research firm based in Waltham, Massachusetts, only one in twenty-five obese people in the United States have prescriptions for drug treatment.

Existing weight-loss drugs have well-known side effects, including faecal incontinence and high blood pressure. They are only mildly effective; none produces an average weight loss of more than 4.5 kilograms. So many insurance plans will not reimburse patients for them, and many patients discard them after a few months. Thousands are turning instead to surgery to lose weight (see graph, left).

Bernice Welles, an endocrinologist and vice-president at DiObex, a San Francisco-based biotechnology company, told an obesity meeting in Washington DC earlier this month that, if it lives up to its promise, rimonabant will transform this picture.

"Sanofi-Aventis definitely has a blockbuster drug on its hands," says Donny Wong, a biochemist and analyst at Decision Resources. Wong's firm is predicting an annual obesity drug market of \$2.3 billion by 2013, with rimonabant accounting for some 60% of those sales. Other analysts have suggested that sales could go higher than that — but the drug's prospects are dampened by concerns that insurers may refuse to pay for it.

And safety concerns will lurk in the background until the drug is tried and tested in the market. In a letter to *The Lancet* in July, Bernard Hirschel, an infectious disease specialist at Geneva University Hospital, suggested that Sanofi-Aventis test its drug candidate in high-risk groups before it goes on sale (*Lancet* 366, 369; 2005).

Jeffrey Bland, a biochemist and president of Metagenics, a maker of nutritional products based in San Clemente, California, says that diet drugs that act on the brain have had a troubled history. A case in point is the fenfluramine-phentermine combination popularly called 'fen-phen', made by Wyeth of Madison, New Jersey. Its fenfluramine component suppressed appetite by boosting serotonin levels in the brain. But it was found to cause serious damage to heart valves and was withdrawn in 1997. "In central nervous system-mediated medication, we almost always learn something once those products are released that we didn't know before," Bland says.

Others, like Lilly's Caro, predict that in the long term, pharmacological approaches to obesity are unlikely to hinge on any single drug, however effective. They will probably involve several, deployed in combinations devised for each patient. In the short term, however, rimonabant could soon have the obesity medicine chest largely to itself. ■

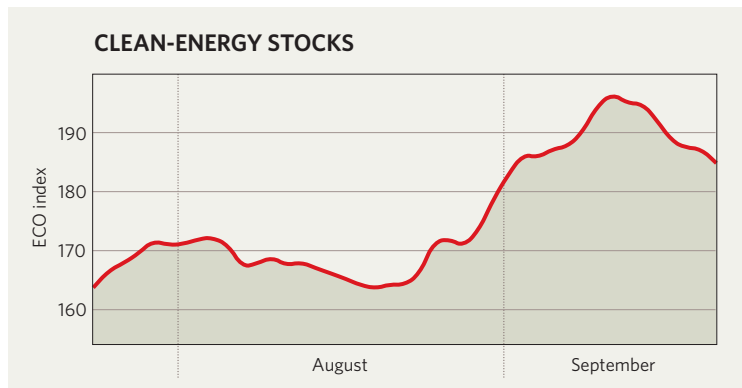
IN BRIEF

UP IN THE AIR The European Commission has introduced legislation that would set standards, for the first time, on pollutants released by cars. The commission says its proposals would reduce premature deaths caused by air pollution from 370,000 to 230,000 each year by 2020 — and cost €7 billion (US\$8.5 billion) annually to implement. Until now, European regulations have focused on fuel economy standards, encouraging the manufacture of diesel-powered cars that use less fuel, but release particulates into the air. Environmental groups had pushed for even tougher and more expensive legislation, and criticized the proposed rules.

PAYOUT TIME GlaxoSmithKline will pay the US Department of Justice \$150 million to settle a claim that the drug company inflated the prices paid by government healthcare programmes for its anti-nausea drugs Zofran (ondansetron) and Kytril (granisetron). The payout arose after a small healthcare provider informed the government of the alleged overcharging, under a law that allows whistle-blowers to gain financially from a settlement. Justice department officials said that 150 similar cases involving drug firms are under investigation. GlaxoSmithKline admitted no wrongdoing in the settlement.

FLU READY The US health department has given a \$100-million contract to Sanofi-Pasteur to manufacture an avian flu vaccine that has just recently proved safe and effective in human volunteers. The vaccine against H5N1 avian influenza — which is threatening to become a global pandemic — will be produced this month and next at the company's US headquarters in Swiftwater, Pennsylvania. It will contribute a yet-to-be-fixed number of doses to a stockpile of 20 million vaccines that the government hopes to build up. The vaccine producer is a division of the French drugmaker Sanofi-Aventis.

MARKET WATCH



Alternative energy is back in vogue — in the most unfortunate of circumstances. With crude oil and gasoline prices at record levels, investors who once dismissed 'clean' energy as a backwater are changing their minds.

The WilderHill Clean Energy Index — whose symbol on the American Stock Exchange is ECO — tracks energy companies that have alternative-energy interests. It has moved sharply upwards since Hurricane Katrina began to threaten Louisiana's oil installations in late August.

Even more significant, according to Robert Wilder, a former political scientist at the University of California, Santa Barbara, whose company runs the index, is the influx of cash into a fund he has created that tracks its performance. About \$100 million has flooded into the fund in the past month, Wilder says — ten times the rate of investment it enjoyed earlier in the year.

"This sector is getting hot," he says. "When oil hits \$70 a barrel, people start thinking about alternatives."

The value of the index is now approaching its 2001 market peak. But this time, Wilder says, "it is much less speculative" than during the dot-com boom, when some stocks were trading at price-to-earnings ratios of up to 100. Now clean-energy stocks have ratios in a more sensible range, typically in the low twenties.

Analysts warn that worries about oil supply are likely to persist, as exporters may get used to prices of \$50 a barrel or more. But Wilder warns clean-energy investors not to get carried away. "People tend to buy a product after it has gone up," he says. "I wouldn't be surprised to see this index go down." And last week, even as Hurricane Rita threatened Texan oil refineries, the index slipped back on profit-taking. ■

SOURCE: WILDERSHARES

Small groups find fatal purpose through the web

SIR — Analyses reported in your News story “Psychologists warn of more suicide attacks in the wake of London bombs” (*Nature* **436**, 308–309; 2005) depict suicide terrorism as the result of organized campaigns aimed at achieving clear political goals, such as national liberation.

These analyses come from studies of conflicts in areas such as the West Bank and Chechnya, which, although important, may not be applicable to recent attacks. Our research leads us to believe that small-group dynamics and values can trump rational self-interest to produce horrific behaviour in ordinary people.

Bruce Hoffman, of the RAND Center for Terrorism Risk Management Policy in Washington DC, finds that 81% of suicide attacks since 1968 occurred after the terror attacks of 11 September 2001, with 31 of the 35 groups held responsible being Islamic militants or ‘jihadi’. Independent studies by the Nixon Center think-tank and by former US intelligence officer Marc Sageman (presented to the World Federation of Scientists Permanent Monitoring Panel on Terrorism in Sicily, May 2005) reveal that more than 80% of known jihadis live in diaspora communities, often marginalized from the host society, and in hard-to-penetrate social networks that consist of about 70% friends and 20% family. Seeking a sense of community, these small groups bond as they surf jihadi websites to find direction and purpose. In the past five years alone, jihadi websites have increased in number from fewer than 20 to more than 4,000.

“Insights into home-grown jihadi attacks must come from understanding group dynamics and psychological motivations.”

— Scott Atran, Jessica Stern

European jihadis act, not to achieve a clearly specified political goal, but to oppose a perceived global evil. Reuven Paz, former research director for Israeli intelligence, reports that even in Iraq, jihadis from 14 other Arab countries say that they have volunteered to fight against ‘international evil’ rather than for Iraq itself (see www.e-prism.org).

From interviewing would-be suicide bombers and sponsors from Europe to southeast Asia, we have learned that terrorism thrives in people who feel humiliated, either in their own lives or through identifying with others, as seen, for example, in reports from Abu Ghraib prison. We ask questions such as: “What if your family were to be killed in retaliation for your

action?” Almost all answer that, although they have a duty to their families, their duty to God comes first. “And what if your action resulted in no one’s death but your own?” The typical response is “God loves you the same”. Such reasoning is not very sensitive to standard cost–benefit calculations or moral trade-offs.

How do we deal with this decentralized global jihadi community? Insights into home-grown jihadi attacks must come from understanding small-group dynamics and psychological motivations, including those that are religiously inspired.

Given the increasing role played by the Internet, efforts should foster alternative peer groups in cities and cyberspace, showing the same commitment and compassion towards their own members as terror groups seem to offer, but in life-enhancing ways and also towards others.

Scott Atran*, Jessica Stern†

*CNRS, Institut Jean Nicod, 1 bis Avenue Lowendal, 75007 Paris, France and Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48106-1248, USA
†Kennedy School of Government, Harvard University, Cambridge, Massachusetts 02138, USA

Most radiation-related deaths happened in 1945

SIR — The figures given in your News story “Shadow hangs over research into Japan’s bomb victims” (*Nature* **436**, 610–611; 2005) are not backed by research carried out at the Radiation Effects Research Foundation (RERF) in Japan.

The atomic bombings killed an estimated 120,000 people in Hiroshima and another 70,000 in Nagasaki by late 1945, including those who died of radiation sickness in the weeks after the bombs were dropped. The number of subsequent deaths from radiation is much smaller.

Studies carried out at RERF indicate that 94 leukaemia deaths have been attributed to radiation exposure since RERF’s recording began in 1950, and 477 radiation-related deaths from solid cancers (D. L. Preston *et al. Radiat. Res.* **162**, 377–389; 2004). Some further radiation-associated deaths may still occur. The RERF cohort (120,000 initially, of whom 43% are still alive) comprises about half of all survivors, but includes almost all who received the highest exposures.

The risk of death from radiation-related disease other than cancer is much lower. Because of higher background non-cancer mortality, the number of non-cancer deaths attributed to radiation is estimated to be about 40% of the number of radiation-related cancer deaths. This brings to about 800 the total number of deaths since 1950 that we

can relate to radiation from the atomic bombs, with perhaps an equal number of radiation-caused deaths yet to occur.

With regard to other points in your story, I would like to mention that RERF’s buildings are sturdy and well maintained and that I have confidence in the pledges made by the Japanese and US governments to continue funding this most important study of radiation effects and risks. A further five-year funding agreement will be signed in November.

Burton Bennett

Radiation Effects Research Foundation,
5-2 Hijiyama Park, Minami-ku,
Hiroshima City, 732-0815 Japan

Public disclosure could deter conflicts of interest

SIR — Your Business story “Fears rise over leaks of clinical trial results” (*Nature* **437**, 191; 2005) describes a conflict-of-interest scandal in which US medical researchers with inside knowledge of ongoing clinical trials are being paid for information they provide as consultants to Wall Street analysts and investors.

The National Institutes of Health (NIH) could start fixing this problem, at least for its own grantees. NIH-funded researchers are required to provide details of any consulting arrangements to their universities, which in turn approve or veto the plans. This information is confidential and usually cannot be seen by the public.

The NIH could require grantees to make public disclosures of their paid arrangements with pharmaceutical, investment and other companies, as well as their ownership of stock and stock options, as a condition of having their medical research funded by the government. The private finances of any US senator or representative can be checked in an instant through links at www.opensecrets.org/pfds. Why not create, by law, a similar system for medical researchers who receive government funding?

A proposal to require readily accessible financial disclosure will probably be fought tooth and nail by those who benefit from leaving things as they are: some university researchers and administrators, officials at the NIH and scientists in industry.

It is an inescapable fact, however, that the partnership of academia, government and industry is plagued by unseen practices that are ethically or legally suspect.

One way to attack this problem is through a requirement for financial disclosure that the public can see.

Ned Feder

National Institutes of Health,
Two Democracy Plaza, Bethesda,
Maryland 20817, USA

BOOKS & ARTS

Diversity and controversy

Why did a well-intentioned effort to understand human evolution go so wrong?

Race to the Finish: Identity and Governance in an Age of Genomics

by Jenny Reardon

Princeton University Press: 2005. 312 pp.
\$55, £35.95 (hbk)/\$17.95, £11.95 (pbk)**Diane Paul**

The Human Genome Diversity Project had a short and troubled life. The aim was to sample and preserve DNA from “isolated indigenous populations” before social changes rendered them useless for the purpose of answering questions about human evolution. But from its birth around 1991 to its unofficial death less than a decade later, indigenous-rights groups attacked the project as racist and neocolonialist, branding it the ‘Vampire Project’. The effort ultimately became an embarrassment to its funders. Today, research on human genetic variation flourishes, but under other rubrics and largely under the radar of Diversity Project critics.

As Jenny Reardon stresses in her book *Race to the Finish*, the project’s leaders were well-intentioned and had impeccable anti-racist credentials. So why did their effort draw unrelenting hostility from groups representing indigenous peoples, some physical anthropologists and others? And could the critics’ fears have been allayed without gutting the project?

As Reardon tells it in this engrossing and even-handed book, the scientists never knew what hit them, and so were unable to mount a response to the project’s detractors. The scientists involved believed they were in a race against time to answer compelling questions about human origins and migrations. But the peoples on whose cooperation the project depended — or at least those claiming to speak for them — were not interested in the scientists’ questions about human origins (to which they already had satisfying answers), disliked being thought of as a resource, took umbrage at the assumption that they were vanishing, mistrusted the project leaders’ motives, especially in regard to patent issues and, in general, did not see what was in it for them.

The organizers struggled to comprehend this reaction. A politically progressive and socially sensitive lot, they knew they were not out to make money, but to pursue what in their view was important and urgent research. To be tarred with the brush of racism — especially

Researchers faced opposition to their plans to understand the origin of indigenous peoples.

given their personal histories — must have been galling. Luigi Luca Cavalli-Sforza had been a trenchant critic of William Shockley’s claim of black genetic inferiority; Robert Cook-Deegan had a long record of involvement with Physicians for Human Rights; and Mary Claire King had worked with the grandmothers of the Plaza de Mayo to identify children kidnapped during Argentina’s dirty war. But avowals of their good intentions did not mollify critics, and organizers eventually set about addressing specific concerns.

Unfortunately, in trying to solve one problem, they often created another. For example, in addressing concerns over whether subjects could give truly informed consent, organizers drafted a model ethical protocol incorporating the innovative concept of group consent. But, in a particularly rich chapter, Reardon shows that instead of alleviating worries, it ultimately prompted new concerns about paternalism and the revival of old colonialist and racist categories.

In trying to respond to criticism and build legitimacy for the project, the evolutionary biologists and population geneticists who launched it constantly widened the circle of those consulted. Anthropologists and bioethicists were brought into the fold. But those who felt excluded wanted to speak for themselves.

In time, indigenous-rights organizations, African-Americans and Native Americans were also invited to join the discussion, raising some thorny questions about the identity of groups and who was authorized to speak for them.

In any case, these groups were themselves divided and so were impossible to enrol as a unit. Thus there were both supportive and critical voices within anthropology — the most fractured of disciplines — and although some African-Americans and Native Americans were attracted to opportunities offered by the project, others feared co-option and saw efforts to include them as bribes. For some opponents, to even critique a proposal would grant it legitimacy.

Could the project have been saved? Reardon believes that it might have been had discussion gone much deeper, with sustained attention to questions of the nature of scientific knowledge and its relation to power. It seems that the moral of the story is the need to include scholars from the field of science studies, who could have introduced a more sophisticated framework for thinking about race and power in genetic research. Perhaps. But it may be that even then a solution that satisfied critics while preserving the project’s core was simply unachievable.

In the event, the critics stopped the project

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

in its tracks. Reardon sees little to celebrate in this victory. The project's proponents correctly predicted from the start that, if they failed, the research would continue but in a much less public and organized way. The study of human genetic variation is now fashionable, but it is being pursued without scrutiny of the deeper issues that Reardon believes essential to the pursuit of both a more reflective science and a

more sensitive society. Funders have understandably tried to avoid the controversies that sank the Diversity Project. But the ironic result has been to narrow discussion of the issues at stake even further.

Diane Paul is in the Department of Political Science, University of Massachusetts, 100 Morrissey Boulevard, Boston, Massachusetts 02125-3393, USA.

Touching a nerve

The War of the Soups and the Sparks: The Discovery of Neurotransmitters and the Dispute over How Nerves Communicate
by Elliot S. Valenstein

Columbia University Press: 2005. 256 pp.
\$31, £19.50

Charles Stevens

Controversy is an inevitable, and essential, part of science, but one that scientists generally find uncomfortable and tend to regard as a blemish to be hidden from the public. Elliot Valenstein's book *The War of the Soups and the Sparks* is a readable and instructive history of one of neuroscience's most important scientific disputes, the three-decade debate about how neurons communicate with one another. He explains the way our current views developed and places the work in its social and human context by providing biographical sketches that bring the participants to life.

One neuron sends information to another at a point of contact known as a synapse. We now understand this process of information transfer in great detail, and know that it involves the release of a chemical, called a neurotransmitter. Valenstein's book is about how we arrived at this picture, and especially about the controversies along the way. The main debate related to whether synaptic transmission is chemical — that is, whether the information-carrying signal is the release of a neurotransmitter — or whether it, like the nerve impulse itself, uses purely electrical signals.

The story begins about a hundred years ago with investigations of how nerves influence the function of organs; an example is the slower heart beat produced by stimulating the vagus nerve. By 1920 it was firmly established, largely by Henry Dale, that acetylcholine, a chemical not known at the time to occur in the body, also decreased the heart rate and duplicated various effects of nerve stimulation on other organs. But the idea that the vagus nerve secreted acetylcholine or something similar was not considered: nerves are tiny, seemingly too small to be the source of hormone-like chemicals.

Starting in 1921, Otto Loewi published a series of papers claiming that the vagus nerve secretes some chemical — he called it *Vagusstoff* — when stimulated, which slows

the heart. Loewi's work met with great scepticism, partly because he could not show that the *Vagusstoff* came from the vagus nerve and not the heart, but mainly because others could not repeat his technically tricky experiments. Using improved techniques, Dale identified the *Vagusstoff* as acetylcholine and showed that it was released by the stimulation of many different nerves that affect the function of various organs. By 1936 the conclusion that neurotransmitter is released at synapses outside the brain was well enough established to attract a Nobel Prize for Loewi and Dale.

In 1936, physiologists could accept that the neurotransmitter released by the vagus nerve slows the heart and has other slowly developing effects. But they could not believe that this mechanism could cause rapid events such as the contraction of skeletal muscle or communication in brain circuits. Instead, they were convinced that this must be due to a direct spread of current from nerve impulses, known to involve electrical rather than chemical signals.

Almost all neurophysiologists believed that synaptic transmission had to be electrical, rather than chemical. One of the most prominent opponents of chemical transmission for fast synapses was John Eccles, a friend and great admirer of Dale. Their debate was vigorous, but good-natured and respectful. Valenstein points out that the physiologists believed that only electrical transmission could be fast enough, but also that the dispute was a class war between pharmacology and physiology. The physiologists used modern, sophisticated methodology and tools, such as the cathode-ray oscilloscope, whereas pharmacologists were still using bioassays, such as leech muscle, and old-fashioned recording methods. The physiologists looked down on the pharmacologists, and felt that conclusions based on methods less sophisticated than their own were not to be trusted.

This dispute continued until the middle of the twentieth century, when results from new technology finally convinced the physiologists that synapses do communicate by the release of neurotransmitters. Eccles, one of the strongest proponents of electrical transmission at synapses, provided some of the key evidence showing that he, and the other physiologists, had been wrong.

Why did this argument last so long? As with all such disputes, part of the reason was that technology was not available that could provide decisive tests of the alternative possibilities; the correct answer came with technological advances. The other reason is that synaptic transmission is much more complex than either side envisaged, and the discussion was framed in simplistic terms because the scientists involved sought simplicity where it did not exist.

Charles Stevens is at the Salk Institute for Biological Studies, PO Box 85800, San Diego, California 92186-5800, USA.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

A stimulating debate: Henry Dale (top) and Otto Loewi showed that nerves release chemicals.

Roving the Solar System

The Planets

by Dava Sobel

Fourth Estate/Viking: 2005. 288 pp.
£15/\$24.95

William K. Hartmann

Dava Sobel, the science reporter known for her brilliant books *Longitude* and *Galileo's Daughter*, now gives us a beautifully written rumination about the planets and small bodies of the Solar System. In *The Planets* she starts with Mercury and works outwards from the Sun. Those who expect an overview of modern planetary science will be disappointed, however. This modest-sized, pleasant book is a master raconteur's meander through history and astrology, and it dips into modern discoveries only sporadically. The late Carl Sagan covered some of the same territory in several of his books, but, perhaps because he was as an active researcher, he gave more of a sense of the mysteries, discoveries and ultimate consequence of cosmic exploration.

I'm a fan of Sobel's historical and cultural sensibilities, and I believe in teaching the history of science, but I get the feeling that someone's concept of a literary book about planets gets in the way here. For example, there's an odd patchwork scheme with each chapter being in a different style. The Earth chapter is virtually all history, yet written in the present tense: "Darwin is sailing" and so on. The Mars chapter, oddly titled "Sci-fi", is a first-person narrative by a 4.5-billion-year-old martian meteorite. The chapter on Uranus and Neptune is mostly in the form of an imaginary letter from the British astronomer Caroline Herschel to her American counterpart Maria Mitchell. A final chapter dwells on a swell party at researcher Andy Ingersoll's house

after the Cassini probe reached Saturn.

Sobel's great strength appears in the chapter on Earth. Here's a real story — the tale of the Earth emerging in human minds as a planet. Here's Ptolemy, recognizing that timing lunar eclipses from different cities would allow estimates of longitudes. Here's Gilbert, discovering magnetism in 1600 and venturing a pre-newtonian view that it might be the force that keeps planets in their orbits. And here's Halley, urging that observers after his death

watch for his predicted Venus transit, in order to triangulate the interplanetary distance scale. This beautiful chapter reminds us that scientists are now in the midst of a progressive

adventure, something that the public and most journalists fail to grasp — especially as American fundamentalists recycle erroneous seventeenth-century arguments about the age of Earth, arguments long since settled in Europe.

My qualms returned when I found that many significant physical concepts, such as hydrogen fusion and orbital resonances, are consigned to the oblivion of a 14-page small-print "Details" section, along with additional mythological and historical tidbits. Did some editor win an argument that the delicate sensitivities of readers should not be troubled with pesky facts and basic principles? Moreover, as a colleague pointed out, this section erroneously states that most 'tidally bound' moons are in a 2:1 spin resonance, unlike our Moon. In fact most moons, like ours, are in synchronous 1:1 rotation.

I must also mention my disquiet when the review copy I received was not a marketable copy but an advance review copy. Books, like many other products, are beginning to offer a sobering case study of the so-called free market. The number of distributors providing books to stores has collapsed to just a few, and several are owned by the big bookstore chains. When these distributors make mass purchases to place a book in their own stores, then a book's readership, like an election's outcome, ends up depending on which products are most aggressively marketed. With the marketing departments in control, orchestrated campaigns are the name of the game. Reviewers in such a situation can become part of the advance promotional machinery. To be fair, the final hardback release arrived just as I was about to send in my review. Typos seemed to have been corrected, but it confirmed that the book ignores modern spacecraft imagery.

It is a nice book and it might be a good gift for a literately inclined or 'artsy' friend who is dubious about science. I was taken aback, however, that in an age of rovers on Mars and landers on Titan, a book named *The Planets* offers the public mostly astrology, mythology, history and archival woodcuts, interesting though they may be. Somehow it reminded me of the genius of the US political consultant Karl Rove, who manages to keep people enthused by distracting them from the real issues.

William K. Hartmann is at the Planetary Science Institute, 1700 East Fort Lowell Road, Suite 106, Tucson, Arizona 85719, USA.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Virtual life

Louis Bec, a member of the 'artificial-life art' movement, uses computer models to artificially evolve new virtual species from existing organisms. His fabulous zoomorphic forms include the *Melaskunodousse* shown here, which is evolved from several generic ancestors. The French biologist styles himself as the only zoosystematician in the world.

This computer image is one of nearly 300 works described in the book *Kunst aus dem Labor* [Art From the Laboratory] by Ingeborg Reichle (Springer, €49). The artworks analysed by Reichle range from Salvador Dali's *Butterfly Landscape* to Suzanne Anker's contemporary installations, which resonate so powerfully with genetics research.



NASA/JPL

LOUIS BEC

Thinking big

Fritz London's single-minded thinking led him to surpass even Einstein, as he believed correctly that quantum mechanics was right at all scales, including the macroscopic.

Philip W. Anderson

Fritz London began his career in physics as one of the originators of quantum theory during 1925–27. His training as a philosopher, before taking up physics, no doubt enhanced his contribution to the ‘copenhagen interpretation’ — the first general attempt to understand the world of atoms according to quantum mechanics. But London did much more than create the first theory of the chemical bond, and has not had the recognition he deserves.

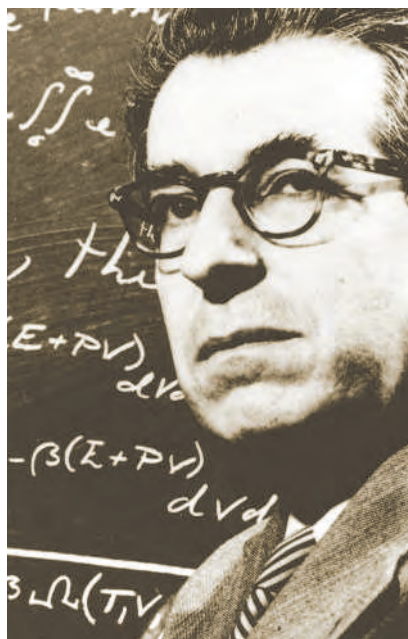
He was among the few pioneers who deliberately chose, once atoms and molecules were understood, not to focus his research on further subdividing the atom into its ultimate constituents, but on exploring how quantum theory could work, and be observed, on the macroscopic scale.

For a few years, London worked at trying to found chemistry on quantum theory, but in the end was overwhelmed by Linus Pauling's more heuristic approach; he never published his book on the subject. He then became intrigued by the twin phenomena of superfluidity and superconductivity, which, he was convinced, were macroscopic manifestations of quantum mechanics.

In 1935, London was the first to propose that superfluidity was Bose–Einstein condensation, and then in the late 1930s, with his brother Heinz, he developed the first heuristic theory of superconductivity. His pair of books on these subjects appeared around 1950 and admirably framed the questions that were soon to be answered — in the one case by Oliver Penrose, Lars Onsager and Richard Feynman, and in the other by John Bardeen, Leon Cooper and Robert Schrieffer. But London fell ill in 1950 and died in 1954, so he did not live to see the triumphs of his intuitions.

He had paid, however, for his unpopular choice of subject matter — quantum theory on the macroscopic scale — by having to settle for a job in the pre-war South. This meant being out of mainstream physics, and may have resulted in him being excluded from the Manhattan bomb project on which all his early associates worked.

In 1939, in an obscure paper called ‘The observation problem in quantum mechanics’, London and Edmond Bauer took on the notorious Bohr–Einstein debates. This is the earliest paper I know of that expresses the most common-sense approach to the uncertainty principle and the philosophy of quantum measurement.



Lone thinker: Fritz London took an opposite tack from both Albert Einstein and Niels Bohr.

In reading about these debates I have the sensation of being a small boy who spots not one, but two undressed emperors. Niels Bohr's ‘complementarity principle’ — that there are two incompatible but equally correct ways of looking at things — was merely a way of using his prestige to promulgate a dubious philosophical view that would keep physicists working with the wonderful apparatus of quantum theory. Albert Einstein comes off a little better because he at least saw that what Bohr had to say was philosophically nonsense. But Einstein's greatest mistake was that he assumed that Bohr was right — that there is no alternative to complementarity and therefore that quantum mechanics must be wrong. This was a far greater mistake, as we now know, than the cosmological constant.

At this point London took an opposite tack from either Bohr or Einstein. He found it difficult to believe Bohr's idea that there was a real ‘complementarity’ even though he had been an early contributor to that line of thinking. Instead he took the then radical step of assuming that quantum mechanics was not wrong, but right at all scales, including the macroscopic. This explains why London was intrigued by the realization that in the ‘super’ forms of matter, he was seeing quantum theory showing itself on the (relatively) everyday scale.

Taking London's point of view, one

immediately begins to realize that the real problem of quantum measurement is not in understanding the simple electron that is being measured, but the large and complicated apparatus used to measure it. This apparatus has all kinds of properties that are not obvious consequences of quantum mechanics: rigid slits, for instance, and a photographic plate that darkens irreversibly where an electron hits it.

These properties are a real intellectual challenge to understand from first principles; the first thing one realizes is that time, for the measurer and the photographic plate, has a sign — earlier or later. This sign is not contained in the quantum theory and has to be the result of the organizing principles of quantum particles assembled into very large macroscopic objects. This and the fact that the apparatus has a definite position in space require that a quantum description of it can only be given in terms of a superposition of an unimaginably large number of different quantum states.

The electron interacting with it attaches (entangles) one part of its wave function to one batch of these states, the other part to a different batch. And these batches differ in so many ways that they can never be made to cohere again; they represent two entirely separate macroscopic histories of the apparatus. The message is that what is needed is an understanding of the macroscopic world in terms of quantum mechanics. This is the direction that London chose.

And that brings me to superfluid solids. Moses Chan and his student Eun-Seong Kim have recently shown that helium (and probably hydrogen), if solidified below a tenth of a degree Kelvin, flow through their own crystal lattice like a superfluid. (This has yet to be confirmed, but I believe it.) This means that a rigid object — the most primitive of our physical intuitions — is not a system in a simple, single quantum-mechanical ground state, but only arises as a consequence of thermal fluctuations.

Thus, Albert Einstein's clocks and rigid measuring rods, which play such a key role in the theory of relativity, must be not primitive but derived in a very complex way from the underlying quantum laws of microscopic physics. At which point I could immodestly take the opportunity to announce that after all, “more is different!”

Philip W. Anderson is in the Department of Physics, Princeton University, Princeton, New Jersey 08544, USA.

E. SEGRE VISUAL ARCHIVES/PHYSICS TODAY COLLECTION

ESSAY

NEWS & VIEWS

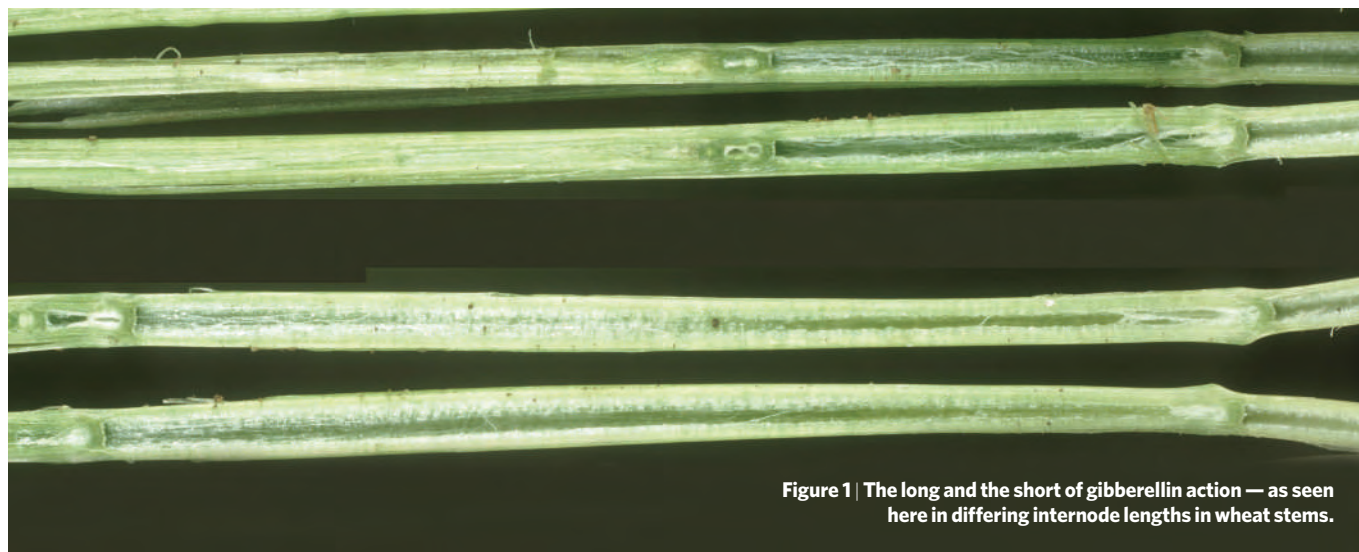


Figure 1 | The long and the short of gibberellin action — as seen here in differing internode lengths in wheat stems.

PLANT BIOLOGY

A receptor for gibberellin

Dario Bonetta and Peter McCourt

The identification of a receptor for gibberellin, a plant signalling molecule, opens up new prospects for understanding plant growth and development. Not least, crop-selection programmes should benefit.

Plants have an astounding ability to respond to external conditions — consider the practice of bonsai, in which a potentially mighty tree can be duped into becoming a potted plant. This developmental plasticity has been manipulated throughout agricultural history, most notably during the 'green revolution' of the past century¹ in which plant breeders often doubled grain production by selecting for semi-dwarf varieties of wheat and rice. The underlying developmental mechanisms responsible have been unclear. But the synthesis of (or sensitivity to) a hormone called gibberellin, which influences processes usually related to cellular expansion, was clearly involved². Hence the ensuing race to understand how gibberellin is produced, perceived and interpreted by the plant. On page 693 of this issue, Ueguchi-Tanaka *et al.*³ describe how they have identified a receptor for gibberellin, and so reveal a mechanism through which the hormone is perceived.

The initial framework for this research was a set of mutants deficient in gibberellin synthesis. Their characteristics indicated a role for gibberellin in processes ranging from seed germination and leaf expansion to flowering and stem elongation (Fig. 1). With these mutants available, new lines of inquiry could be formulated. Through genetic analysis in

model plant systems, particularly *Arabidopsis*, it was recognized that a group of repressor proteins are at the core of gibberellin signalling^{4–6}. Designated DELLA domain proteins because they share a short amino-acid sequence, these proteins repress a variety of downstream targets, which include gene transcription factors that promote gibberellin-related processes.

The addition of gibberellin releases this repression by somehow causing DELLA proteins to be degraded⁷ (Fig. 2, overleaf). Genetic analysis in both *Arabidopsis* and rice identified part of an enzyme, the F-box subunit of an SCF E3 ubiquitin ligase, as a cog in this degradation mechanism^{8–10}. Because DELLA proteins are so central in gibberellin signalling, it is perhaps not surprising that one of these DELLA genes, *Rht*, turned out to be the magic bullet that led to the semi-dwarf wheat varieties of the green revolution¹¹.

Although these findings closed the loop between basic research and agricultural breeding, the missing player has been the gibberellin receptor. It has long been postulated that plant cells have both membrane-bound and soluble receptors, but it is not known how many might exist. Moreover, there could be cell-specific receptors or ones specific for the isoforms of gibberellin typically found

in plants. Although the DELLA protein-degradation pathway needs to be activated by gibberellin to cause protein destruction, none of the components of this pathway bind gibberellin and so are not themselves receptors.

Ueguchi-Tanaka *et al.*³ concentrated on a gene called *GID1*, which they identified through the map-based cloning of a rice mutant. Various lines of evidence suggest that this gene encodes a gibberellin receptor. First, variants of *GID1* that cause loss of function produce plants that cannot respond to gibberellin, indicating that GID1 protein acts as a positive regulator of gibberellin signalling. Second, through double-mutant analysis the authors show that *GID1* acts at or upstream of a rice DELLA gene, *SLR1*, and that gibberellin-induced SLR1 degradation depends on functional GID1 protein. Fusions of GID1 with a marker, green fluorescent protein, indicate that GID1 is soluble and primarily located in the nucleus, its proposed site of action.

Importantly, the authors go on to show that recombinant GID1 protein can specifically bind radiolabelled gibberellin, and that GID1 binding to SLR1 is gibberellin-dependent. Finally, overexpression of GID1 in transgenic rice results in long, spindly plants, an outcome that would be expected from an overdose of gibberellin.

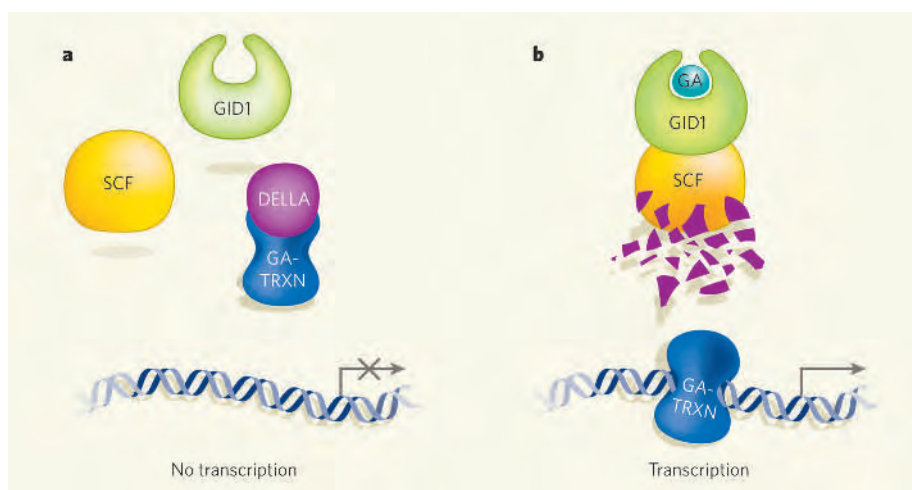


Figure 2 | Gibberellin action in rice. **a**, In the absence of gibberellin, repressor proteins (DELLA) interfere with gibberellin-dependent transcription factors (GA-TRXN). **b**, When gibberellin (GA) binds the GID1 protein, now identified as a receptor³, GID1 can interact with the protein turnover complex (SCF). The SCF complex is then able to degrade the DELLA repressor, thereby freeing GA-TRXN to stimulate gene transcription.

The mechanism of gibberellin perception and action is intriguing, because earlier this year it was revealed that another plant hormone, auxin, similarly acts by stimulating components involved in targeted protein destruction (auxin stimulates interactions between an F-box protein and the rest of the SCF complex^{12,13}). So plants may commonly use small organic molecules to change protein–protein interactions that then go on to direct development. This possibility bears on a long debate about whether plant hormones can be called hormones at all — an issue that goes beyond semantics and influences our understanding of plant signalling mechanisms as well as their evolution.

Many plant hormones do not fit the classic definition of substances that act at a distance, and instead often seem to be made close to where they act. In addition, most plant hormones are small, side-products of metabolism, not large polypeptides. So is it surprising that the mechanisms for plant hormone perception differ from those seen in multicellular animals? Most plants are of course also multicellular. But they are essentially immobile composites of nearly independent identical units, and having centralized control centres that coordinate aspects of physiology or growth does not necessarily make the most sense.

Instead, the different units are likely to experience different local environments and localized responses, tuned to those environments, that allow better overall coordination than if the information was interpreted at a remote location. It seems that plants make up for their lack of morphological complexity with metabolic complexity; metabolic networks are continually being adjusted in response to fluctuating environments. Plant hormones seem to be part and parcel of metabolism, with increased hormone levels often preceding the ultimate response. So the signalling pathways that we are just now discovering, the result of a

long evolutionary history, could be offshoots of complex metabolic networks.

A receptor for gibberellin has been an elusive beast. The successful hunt carried out by Ueguchi-Tanaka *et al.*³ not only takes the plant

hormone debate further, but the enhanced molecular understanding of gibberellin signalling may also presage a new green revolution. At the least, the findings will feed plant biologists' appetites for many years. ■

Dario Bonetta is in the Faculty of Science, University of Ontario Institute of Technology, Oshawa, Ontario L1H 7K4, Canada.

Peter McCourt is in the Department of Botany, University of Toronto, Toronto, Ontario M5S 3B2, Canada.
e-mails: dario.bonetta@uoit.ca;
mccourt@botany.utoronto.ca

1. Davies, W. P. *Nutr. Rev.* **61**, S124–S134 (2003).
2. Boerner, A., Plaschke, J., Korzun, V. & Worland, A. J. *Euphytica* **89**, 69–75 (1996).
3. Ueguchi-Tanaka, M. *et al.* *Nature* **437**, 693–698 (2005).
4. Peng, J. *et al.* *Genes Dev.* **11**, 3194–3205 (1997).
5. Silverstone, A. L., Ciampaglio, C. N. & Sun, T. P. *Plant Cell* **10**, 155–169 (1998).
6. Ikeda, A. *et al.* *Plant Cell* **13**, 999–1010 (2001).
7. Dill, A., Jung, H. S. & Sun, T. P. *Proc Natl Acad. Sci. USA* **98**, 14162–14167 (2001).
8. Sasaki, A. *et al.* *Science* **299**, 1896–1898 (2003).
9. Dill, A., Thomas, S. G., Hu, J., Steber, C. M. & Sun, T. P. *Plant Cell* **16**, 1392–1405 (2004).
10. McGinnis, K. M. *et al.* *Plant Cell* **15**, 1120–1130 (2003).
11. Peng, J. *et al.* *Nature* **400**, 256–261 (1999).
12. Dharmasiri, N., Dharmasiri, S. & Estelle, M. *Nature* **435**, 441–445 (2005).
13. Kepinski, S. & Leyser, O. *Nature* **435**, 446–451 (2005).

OCEANOGRAPHY

Nutrients in remote mode

Marina Lévy

Phytoplankton productivity depends on the replenishment of nutrients in ocean surface waters. An explanation for a region of strikingly low productivity invokes a little-considered aspect of the nutrient cycle.

In the open ocean, so-called 'mode waters' are water masses of uniform density that form at the surface in winter, subduct and then may travel long distances beneath the surface. Palter *et al.*¹ (page 687 of this issue) show that in one circulation system in the North Atlantic, known as the North Atlantic subtropical gyre, this lateral process adversely affects nutrient availability along the mode-water route — and so, by providing only limited amounts of nitrate and phosphate, exerts a remote control on phytoplankton productivity. Along with other work, Palter and colleagues' observations add a new dimension to our growing appreciation of the various controls on marine productivity.

One aspect of the background to these investigations is that, at mid-latitudes in the open ocean, convection caused by seasonal changes in temperature and wind conditions makes the nutrients required for phytoplankton growth available in the sunlit surface waters. One question that exercises oceanographers is the nature of the mechanisms that refill the nutrient reservoir in subsurface waters.

A second aspect is the large-scale distribution of phytoplankton productivity, in this case in the North Atlantic, as measured by levels of surface chlorophyll (Fig. 1). The North Atlantic can be broadly divided into two huge circulation systems. To the north of the Gulf Stream, the subpolar gyre is productive, with high concentrations of chlorophyll. To the south, the subtropical gyre is a vast biological desert. The traditional view is that this large-scale contrast is determined by two factors². One is wind-driven (Ekman) pumping, which induces upwelling of water in the subpolar gyre and downwelling in the subtropical gyre. The other is winter convection, which mixes nutrient-depleted surface waters with nutrient-rich subsurface waters, and which increases towards higher latitudes.

But there are also differing east–west gradients of chlorophyll in the two gyres, which the patterns of Ekman transport or convection^{1,2} cannot explain. There is minimum productivity in the western part of the subtropical gyre (purple region in Fig. 1), and a tongue of maximum productivity in the western subpolar

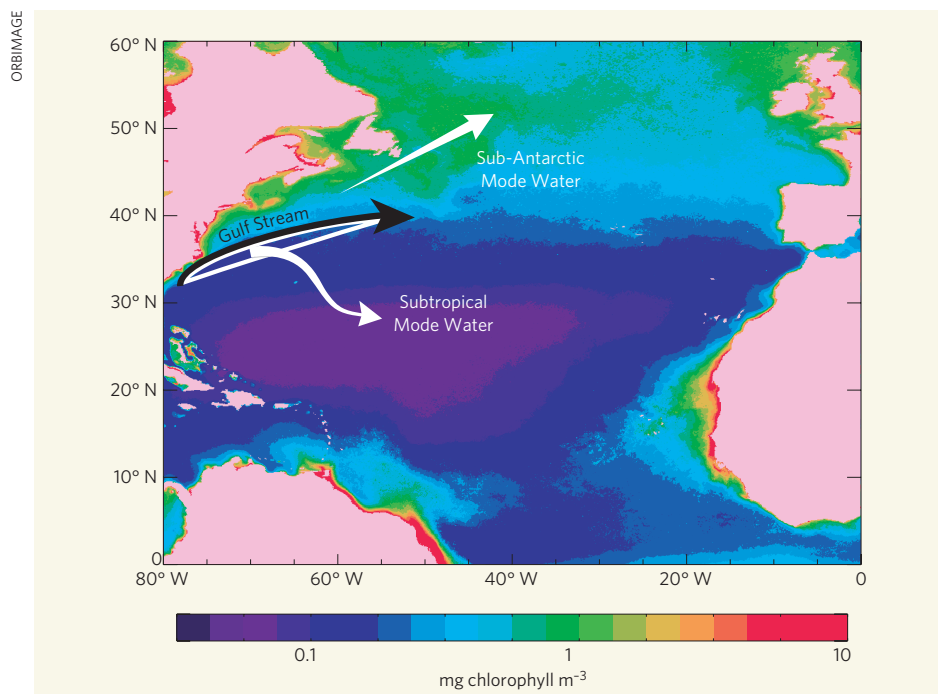


Figure 1 | Distribution of surface chlorophyll in the North Atlantic. Chlorophyll (as seen by the satellite-borne sensor SeaWiFS) is an index of phytoplankton biomass and therefore productivity. The region outlined in white shows where Subtropical Mode Water is formed. Palter and colleagues' explanation¹ of the western productivity minimum in the subtropical gyre is that it stems from the subsurface movement (lower white arrow) of nutrient-depleted Subtropical Mode Water from its site of formation. The western maximum in the subpolar gyre arises, by contrast, from the northern movement of nutrient-rich Sub-Antarctic Mode Water, and its delivery (upper white arrow) at the surface north of the Gulf Stream^{4,5}.

gyre (green region in Fig. 1). This is where nutrient levels in the subsurface reservoir come into the picture.

Palter *et al.*¹ show that the reservoir south of the Gulf Stream is fuelled laterally with Subtropical Mode Water, which forms to the north of the gyre and then circulates southwards beneath the surface. These subsurface water masses are particularly poor in nutrients, as measured in vertical profiles by the World Ocean Circulation Experiment. Palter *et al.* argue convincingly that the chlorophyll minimum in the western part of the subtropical gyre is the signature of this underlying,

nutrient-depleted reservoir when it mixes with the surface waters (Fig. 1).

A likely explanation for the nutrient depletion in Subtropical Mode Water is that phytoplankton growth consumes nutrients on a large scale in winter, at the same time as the mode water is starting to subduct and embark on its subsurface journey. The remineralization process, which replenishes nutrients and occurs at depth, is inadequate to redress this initial nutrient loss on the timescales involved.

But how can winter phytoplankton growth be sustained before the onset of conditions that produce the especially vigorous burst of

growth in spring? The answer may lie in a specific biological regime that pertains over the area of mode-water formation, in which light and nutrient limitations on growth are balanced in such a way that winter growth is greater than it is farther north and farther south. Such a 'mid-latitude regime' is evident in the northeast Atlantic³, in a narrow band between 37° N and 43° N.

Complementary work by Williams *et al.*⁴ fills in the picture in the subpolar gyre. They identify an opposite effect north of the Gulf Stream, in which mode waters are the primary cause of high phytoplankton productivity in the west of the subpolar gyre. In this case, it is the induction flux of nutrients that sustains the high productivity (Fig. 1) — induction is a subsurface-to-surface process and directly provides the sunlit layer with nutrients; subduction, by contrast, is a surface-to-subsurface process that affects the nutrient reservoir. This induction flux⁴ covers a larger area on the western side than on the eastern side of the ocean basin, and so may also explain the east–west gradient in the subpolar gyre.

Using model diagnostics, Williams *et al.* go further, providing evidence that the induction flux is mainly composed of Sub-Antarctic Mode Water⁵, which originates from the Southern Ocean and travels northwards along the western boundary of the Atlantic. This mode water is rich in nutrients, because it is formed in a 'high-nutrient, low-chlorophyll' region where low productivity is the norm.

This is not the end of the story. A further upshot of Palter and colleagues' investigations¹ is the recognition of a source of long-term variations, occurring on timescales of decades, driven by the slow cycle of mode waters. The authors suggest that a drop in the formation rate of Subtropical Mode Water was responsible for the large increase in phytoplankton production in the 1990s, compared with the 1960s, that was observed close to Bermuda (33° 22' N, 64° 41' W) downstream of mode-water formation. This is a counterintuitive proposal. Extended periods of cold winters

FLUID DYNAMICS

Let us spray

The smaller a nozzle is, the faster water at the same initial pressure will spray out — and the smaller the emerging droplets will be. It is a commonplace phenomenon that often surprises, and sometimes delights (see picture). But what happens when the nozzle is very small? And what kind of nozzle produces the smallest droplets?

P. McGuinness and colleagues applied themselves to these questions (*J. Phys. D* **38**, 3382–3386;

2005). Their motivation was by no means a frivolous one: the answers are crucial to improving the resolution of inkjet printing, as well as being more generally applicable to industrial techniques requiring the manipulation of small liquid samples. In such cases, the high surface tension that develops at nozzles of micrometre diameters could limit the scope for reducing droplet size.

So the authors tested different sizes and shapes of small nozzles,



using numerical techniques based on the Young–Laplace equation, which relates the pressure difference at a gas–liquid interface to

its geometry. For two-dimensional (planar) nozzles, a triangular opening with sides curved slightly inwards proved the best choice: compared with a conventional, circular opening at the same pressure, it provided a 16% reduction in droplet volume.

But the authors didn't stop there. By bending the corners of the curvilinear triangle up or down to form a non-planar nozzle tip, they were able to bring the reduction in volume to around 33%. As they point out, this adds another dimension to questions of small-droplet generation.

Richard Webb

(as occurred in the 1960s compared with the 1990s) promote deep convective mixing and vigorous mode-water formation: in a one-dimensional view³, increased convection in the subtropical gyre should lead to stronger phytoplankton growth through the increase in nutrient supply. But Palter *et al.* argue that deep convective mixing in the area of Subtropical Mode Water formation in the 1960s severely diminished primary production in downstream regions because of the ensuing subsurface delivery of nutrient-poor waters.

The past 20 years have seen a great deal of research into how nutrients are delivered to the ocean surface. The variability of the subsurface nutrient reservoir has received much less attention: achieving a better understanding of

that variability is one of the next challenges in marine biogeochemistry. ■

Marina Lévy is in the Laboratoire d'Océanographie et de Climatologie par l'Expérimentation et l'Analyse Numérique, Institut Pierre Simon Laplace, 4 place Jussieu, 75252 Paris cedex 05, France.
e-mail: marina@lodyc.jussieu.fr

1. Palter, J. B., Lozier, M. S. & Barber, R. T. *Nature* **437**, 687–692 (2005).
2. Williams, R. G. & Follows, M. J. in *Ocean Biogeochemistry: A JGOFS Synthesis* (ed. Fasham, M.) 19–50 (Springer, New York, 2003).
3. Lévy, M., Lehahn, Y. *et al.* *J. Geophys. Res.* **110**, doi:10.1029/2004JC002771 (2005).
4. Williams, R. G., Roussenov, V. & Follows, M. J. presented at ASLO Summer Meeting, Santiago de Compostela, 19–24 June 2005; *Glob. Biogeochem. Cycles* (submitted).
5. Sarmiento, J. L. *et al.* *Nature* **427**, 56–60 (2004).

SYSTEMS BIOLOGY

Deviations in mating

Avigdor Eldar and Michael Elowitz

Why do cells of the same type, grown in the same conditions, look and behave so differently? Studying fluctuations in a well-characterized genetic pathway in yeast hints at how such variation arises.

A glance in a microscope quickly convinces one that cells are strikingly diverse. Even when they share the same genome and are grown in the same environment, individual cells differ in size, shape and response to stimuli. Traditionally, such diversity has been a confounding factor in biology experiments, which seek to discover the precise response of a cell to a particular stimulus, but must instead contend with a multitude of answers. Improvements in quantitative single-cell methods and fluorescent imaging techniques now allow researchers to use the inherent diversity among cells to ask two questions: where does variation come from; and what can this variation tell us about the genetic circuits whose underlying components are fluctuating? On page 699 of this issue, Colman-Lerner and colleagues¹ confront these questions, reporting their quantitative single-cell analysis of a classic genetic network: the activation of genes in yeast cells in response to mating pheromone.

The yeast mating pathway has been characterized in great detail by beautiful genetic, biochemical and cell-biological experiments². Yeast cells have two mating types: **a** and **α**. Stimulation of **a**-type yeast cells with **α** factor, the pheromone made by cells of opposite mating type, results in activation of a transcription factor (a gene-regulatory protein) called Ste12p. This protein in turn switches on a set of target genes by interacting with their regulatory sequences (promoters), which include one named P_{PRM1} . Stimulation eventually causes cells to adopt a 'shmoo' shape (a pear

shape with one end elongated) as they seek a mating partner. Pathway output can be tracked by fusing the P_{PRM1} promoter to a 'reporter' gene encoding a fluorescent protein, so that activation of the promoter results in production of the fluorescent protein. All of this makes the mating pathway an ideal system to study at the single-cell level.

Colman-Lerner and colleagues observe that cells vary by about 35% in the amount of gene transcription that occurs in response to **α** factor. So where in the cell does this variation come from? One possibility is the inherent stochastic nature of the biochemical reactions necessary for expression of the reporter gene; this 'intrinsic noise' would be insignificant in a test tube but becomes significant at the small scale of the cell. Alternatively, the gene may be faithfully transmitting fluctuations ('extrinsic noise') in upstream components, such as the transcription factors that regulate it.

One way to discriminate between these two possibilities is to use genes that encode two fluorescent proteins — such as cyan fluorescent protein (CFP) and yellow fluorescent protein (YFP) — that can be distinguished but have the same regulatory sequences. If noise is transmitted from upstream regulatory components then, within each cell, the two colours should be expressed at an equal level, although this level may vary from cell to cell. However, if noise is generated by stochasticity in gene expression, the amounts of the two proteins will differ even within a single cell: they will become uncorrelated. Thus, in general, the level of correlation between the two reporter

genes is determined by the relative significance of the two sources of noise (intrinsic and extrinsic). Such a noise-decomposition experiment directly tests how accurately a cell controls its own gene expression without feedback or higher-level circuitry. Similar assays, applied in yeast and the bacterium *Escherichia coli*, have detected both sources of noise and shown how their relative importance varies with gene, growth condition and expression level^{3,4}.

In yeast, stochasticity in reporter-gene expression, although significant, contributes much less variation than does that from upstream components^{1,4}. Interestingly, previous experiments in yeast showed strong correlations between different promoters, suggesting that variation in one or more cellular components might affect diverse genes in a similar way⁴. Such fluctuations can be attributed to a hypothetical 'global' variable that is correlated with the overall growth rate of the cell and affects many processes, much like the health of a country's economy as a whole affects most stock prices^{1,4,5}. Global fluctuations in gene expression might be related to fluctuations in the concentrations of polymerases or ribosomes — cellular components necessary for the expression of all genes.

Besides the magnitude of variation, the timescale of fluctuations is also important. Colman-Lerner *et al.*¹ report that the responsiveness of a cell to **α** factor is approximately constant over a long timescale of many hours. This is similar to recent observations in *E. coli* showing that the dominant fluctuations are slow, with typical timescales on the order of the cell-cycle time⁵.

The authors attempt to take noise decomposition a step further by introducing a phenomenological model that subdivides transmitted (upstream) noise into two hypothetical categories: global and mating-pathway-specific (Fig. 1a, overleaf). To do this, they built a yeast strain containing the gene encoding YFP under the control of the mating-specific P_{PRM1} promoter and the gene encoding CFP under the control of the pheromone-independent actin promoter, P_{ACT1} (Fig. 1a, top). In such a strain, a high degree of correlation between the two promoters would suggest a common 'global' source of noise; a weak correlation would suggest independent noise sources (Fig. 1a, bottom). Hence, the correlated part is attributed to global noise, whereas the uncorrelated part is attributed to pathway noise.

So where does noise in the mating response come from? The answer is that it depends on the conditions. Experiments performed with different amounts of pheromone show different results (Fig. 1b). At high pheromone concentration, variations in P_{PRM1} expression correspond closely to variations in P_{ACT1} expression. One explanation for this might be that a component of the signalling system has become saturated and hence does not transmit variation in components upstream of it (in fact, Fig. 2b of the paper shows that the mating



50 YEARS AGO

"Man and his machines" — After expressing the fear that technical colleges are not educational institutions but teaching shops, [Mr Harry Rée] emphasized that "in building a world where machines do the work which used to be done by men, it is not good enough to build men who can only work like machines"...

He concluded by envisaging the great contribution educational institutions could make by a counter attack on the creeping disease of passive pleasures which is eating away the soul of modern man. "If we could make the effort...we should look upon automatic factories and computing machines as our benefactors enabling us and our children to taste to the full the real joys of life."

From *Nature* 1 October 1955.

100 YEARS AGO

"The omission of titles of addresses on scientific subjects" — What this busy world wants is help to get at what we are interested in with the least possible waste of time.

This hot haste may seem unbecoming to men of science, or perhaps it may appear that we Americans are in too big a hurry — that we are too much impressed with the motto "time is dollars." But there are many other nimble things we are trying to keep up with, and one of those is the progress of science in Europe, along the lines in which we are especially interested.

If a member of so young and giddy a nation might venture to make a suggestion to older and wiser people, it would be in favour of requesting or requiring the presidents of the various scientific organisations and sections of the British Association to provide headings for their addresses so that those of us who have not the time to read all of these good things may be able at a glance to pick out what we want especially to see.

From *Nature* 28 September 1905.

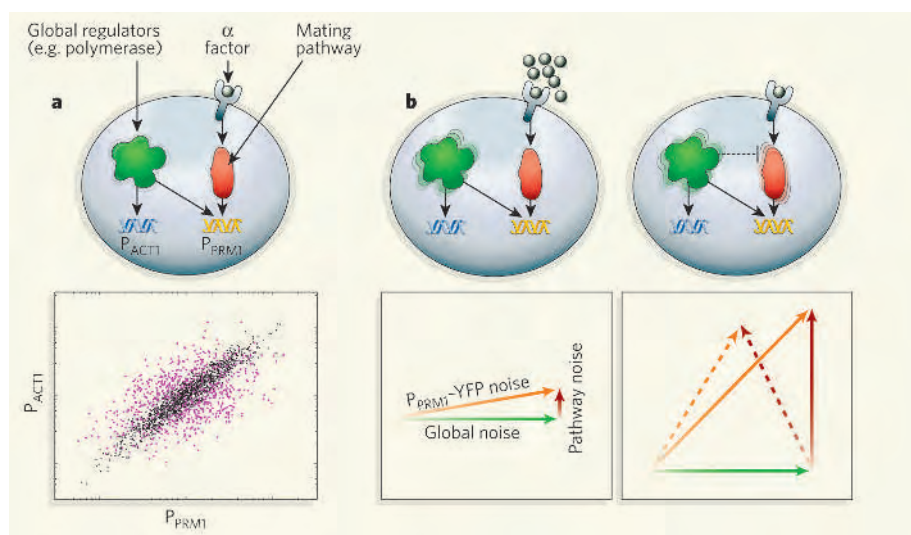


Figure 1 | Noise decomposition in the model of cell behaviour devised by Colman-Lerner and colleagues¹.

a, Upper panel, global regulators (green) affect expression of all genes, including that encoding cyan fluorescent protein regulated by the P_{ACT1} promoter and the gene encoding yellow fluorescent protein (YFP) under the control of P_{PRM1} . The mating pathway (red) affects only P_{PRM1} . Lower panel, expected results in two extreme cases: total P_{PRM1} noise may be dominated by global factors (black points) or by pathway noise (magenta points). **b**, Noise depends on pheromone levels. At high concentrations, noise is dominated by global fluctuations, resulting in strong correlation between the two reporter genes (left). In the vector diagram, coloured arrows represent noise amplitudes and their degrees of correlation with other noise sources. At low pheromone concentration (right), there is reduced correlation between the two reporter genes. Within the model, this is interpreted as an increase in pathway-specific noise. If the two noise sources were independent, the magnitude of P_{PRM1} variation would increase (orthogonal solid arrows, bottom panel). However, the authors observe that the magnitude of the P_{PRM1} variation is independent of pheromone concentration, implying a negative interaction between the two noise sources (acute angle between green and dashed red arrows).

pathway is saturated under these conditions¹). In this case, fluctuations in P_{PRM1} are dominated by the global factors referred to earlier.

But what happens at low pheromone levels? Under these conditions, the authors observed reduced correlation between the two promoters (Fig. 1b, right). If pathway-specific noise were independent of global noise, one would expect the variance in P_{PRM1} activity to be the sum of global and pathway-specific variances (orthogonal green and red arrows in Fig. 1b). This would result in an increased total noise in the P_{PRM1} promoter compared with its value at high pheromone concentration. Interestingly, such an increase is not observed¹. The missing noise indicates the existence of a negative interaction between global factors and the mating pathway (green and non-orthogonal dashed red arrow in Fig. 1b). The authors interpret this negative interaction as a buffering of pathway noise by the effects of global noise on genes in the mating pathway.

The existence of such an interaction is not surprising: being global, such factors should affect other genes along with components of the mating pathway. However, many regulatory steps may connect the global factor to pathway components — the sign and magnitude of this arrow is not clear. It will be interesting to find out whether the inferred buffering interaction has adaptive significance. This could be explored by examining the response of other mating-pathway targets and the effect of variation on the physiological

behaviour of mating. How widespread the buffering is could be examined by similarly analysing other pathways.

In general, correct interpretation of noise experiments can be subtle⁶. For example, the buffering interpretation described above depends on the implicit assumption of the model that global noise affects the P_{PRM1} promoter equally at both pheromone levels (that is, the green arrows in Fig. 1b are equal). In the future, it may be possible to develop higher-resolution analysis of noise correlations, allowing phenomenological models to be replaced by molecular ones. In physics and engineering, analysis of fluctuations often provides unique insights into the dynamics of a system. Normally, such an analysis is limited by the amount of variation present in the system to begin with. Luckily, with living cells, there is no lack of noise to work with. ■

Avigdor Eldar and Michael Elowitz are in the Division of Biology and Department of Applied Physics, California Institute of Technology, Pasadena, California 91125, USA.
e-mail: melowitz@caltech.edu

- Colman-Lerner, A. *et al.* *Nature* **437**, 699–706 (2005).
- Dohlman, H. G. & Thorner, J. W. *Annu. Rev. Biochem.* **70**, 703–754 (2001).
- Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. *Science* **297**, 1183–1186 (2002).
- Raser, J. M. & O'Shea, E. K. *Science* **304**, 1811–1814 (2004).
- Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. *Science* **307**, 1962–1965 (2005).
- Paulsson, J. *Nature* **427**, 415–418 (2004).

MATERIALS SCIENCE

Pore show

Hermann Gies

The holes of mesoporous materials provide sheltered venues for many catalytic and adsorbent processes. A complex and beautiful crystalline germanate structure widens the scope of such materials.

The discovery of ordered mesoporous materials^{1,2} has had an enormous impact on materials research. Such materials have influenced work on gels, surfactants, composite materials, nanomaterials and zeolites (catalytic and adsorbent materials familiar as, for example, water-softeners). Self-organization, surface ('heterogeneous') catalysis and separation technologies have also felt their effects. Writing on page 716 of this issue, Zou *et al.*³ introduce a mesoporous germanium oxide (germanate) material possessing exciting structural properties that could spur further investigations.

Mesoporous materials are three-dimensional bonded atomic networks punctuated by regular nanoscale holes, or pores. (The name 'mesoporous' reflects the size of the holes: materials with significantly smaller pore diameters are known as microporous, those with larger pores as macroporous.) Such materials are generally formed by the action of a negatively charged (anionic) species that include multiple oxygen-containing groups — a silicate, phosphate or aluminate, for example — on a template of a self-assembled organic molecule such as a surfactant. The ionic species match up their charges with those on the surface of the template and so attach themselves to it. The resulting rigid, periodic framework remains stable even when the organic template molecules are burnt away in a process known as calcination, creating a structure possessing pore sizes of 3–500 nanometres.

Mesoporous silicates formed in this way were originally thought to be similar in structure and function to crystalline zeolites possessing smaller pores. This would point to applications in catalysis, separation and sorption technologies — species of high molecular weight could also be used, as these could enter the larger pore spaces and undergo chemical transformations there⁴. But it turns out that using self-assembled surfactant molecules as templates produces amorphous inorganic frameworks: although the structure has mesoscopic order, leading to a characteristic diffraction pattern, atomic-level periodicity in the glass-like silicate framework is lost. This amorphous pore-wall forfeits the shape-selective properties characteristic of zeolites and important for catalysis, acting instead like slightly curved amorphous silica.

This is the remarkable aspect of Zou and colleagues' material³: the germanium oxide

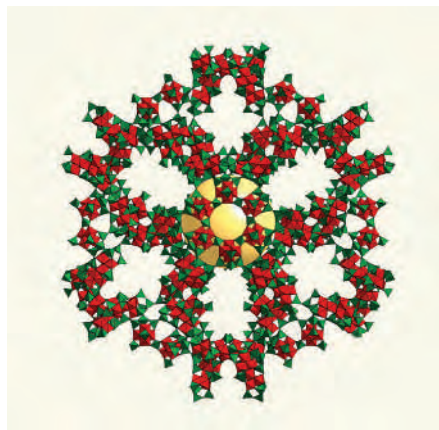


Figure 1 | Mesoporous, yet crystalline. Zou and colleagues' germanate³ combines the regular structure and shape-selectivity of a smaller-pored zeolite with larger pores useful for catalysis. (Green, $[\text{GeO}_4]$ tetrahedra; red, $[\text{GeO}_6]$ octahedra.)

they describe is both mesoporous and crystalline (Fig. 1). It also has the largest cell volume of any inorganic substance, with an accessible pore space that is more than 50% of the total volume. The walls of the structure's pores are formed of germanate clusters with the chemical formula $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$. These are linked through Ge–O–Ge bonds and line a so-called G-minimal surface^{5,6} consisting of a complex series of concave and convex faces with cubic symmetry. (A minimal surface is one that minimizes its surface area — equivalent to its surface energy — for specified boundary conditions, rather like a film of soap in a wire frame used for blowing bubbles.) As the surface of the crystalline walls is significantly curved, the structure contains potential reactive centres that are absent in mesoporous materials lacking crystalline order. This framework fulfils the basic requirement for catalytic activity; its full scope, however, will only become apparent when the germanium sites are replaced by other metal cations.

The periodic G-minimal surface separates a set of two interpenetrating, three-dimensional channel systems that are chiral — that is, non-superimposable mirror images of each other. The interface of these two systems might form the site for stereoselective reactions that produce molecules of a particular spatial configuration. Zou and colleagues also varied³ their synthesis procedure to obtain a material that is chiral overall by blocking off one channel system with a pore filler, leaving

the other channel 'active'. Further work must show whether it is possible to obtain one pure chiral form, and whether the chiral character of the pore can be used in heterogeneous catalysis.

The mesoporous germanate is surprisingly thermally stable, its properties and structure being little affected by heat. The hydroxyl groups that become accessible in its calcined form might also be modified by ion exchange or the introduction of reactive centres into the mesoporous space, providing extra opportunities for catalysis. The smaller pore spaces in crystalline zeolites, and the instability of other ordered mesoporous materials, have until now hampered the successful application of this concept. This potential advance is indicative of a wealth of explorative lines of research using Zou and colleagues' material; time will tell whether all predictions made for its use hold.

The authors have since announced the synthesis of a further crystalline mesoporous germanate, indicating a more general chemical principle behind their findings. In their first structure, the pore geometry is similar to cubic mesoporous structures of a class known as MCM-48, which also have a periodic G-minimal surface. The second material, in contrast, has the pore topology of MCM-41-type material, with a cylindrical minimal surface arranged in a honeycomb pattern.

Zou and colleagues do not speculate on the nature of the physical or chemical mechanisms that allow germanates to yield such highly complex crystal structures. One explanation might be the structure of the $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ anion, with its core of three $[\text{GeO}_6]$ octahedra linked along their edges, shielded by a shell of seven $[\text{GeO}_4]$ tetrahedra linked at their corners. The highly ionic character of the octahedral bond leaves enough flexibility to link subunits, whereas the more directed bond of the tetrahedral units provides the strength and rigidity required to maintain the porous framework, even after calcination.

Unusually, therefore, it is the properties of the building blocks of the pore wall that are central to the formation of Zou and colleagues' germanates. No true templating is required from the template — just space-filling and charge-matching. It will be interesting to see what other building blocks, and which other polyoxo-metallate anions, can be used to construct further such materials. ■

Hermann Gies is at the Institut für Mineralogie, Ruhr-Universität Bochum, Universitätsstraße 150, D-44780 Bochum, Germany.
e-mail: Hermann.Gies@ruhr-uni-bochum.de

1. Kresge, C. T. *et al.* *Nature* **359**, 710–712 (1992).
2. Yanagisawa, T. *et al.* *Bull. Chem. Soc. Jpn* **63**, 988–992 (1990).
3. Zou, X., Conradsson, T., Klingstedt, M., Dadachov, M. S. & O'Keefe, M. *Nature* **437**, 716–719 (2005).
4. Corma, A. *Chem. Rev.* **97**, 2373–2419 (1997).
5. Andersson, S., Hyde, S. T. & von Schnering, H. G. *Z. Kristallogr.* **168**, 1–17 (1984).
6. Lord, E. A. & Mackay, A. L. *Curr. Sci.* **85**, 346–362 (2003).

OBITUARY

Joseph Rotblat 1908–2005

Physicist who committed his life to the cause of nuclear disarmament.

The closing words of Joseph Rotblat's lecture on acceptance of the 1995 Nobel Peace Prize sum up his nature. "The quest for a war-free world has a basic purpose: survival. But if in the process we learn how to achieve it by love rather than by fear, by kindness rather than by compulsion; if in the process we learn to combine the essential with the enjoyable, the expedient with the benevolent, the practical with the beautiful, this will be an extra incentive to embark on this great task. Above all, remember your humanity." Joseph Rotblat died on 31 August, aged 96.

Rotblat was born in Warsaw on 4 November 1908 into a middle-class Jewish family. The family was left impoverished by the First World War: at the age of 15, Rotblat worked as an electrician during the day and studied physics in the evening. He won an open scholarship to the Free University of Poland, later obtaining a doctorate from the University of Warsaw with research on the inelastic scattering of neutrons. It was while working in the radiological laboratory of the Scientific Society of Warsaw that he heard of the discovery of nuclear fission. He then himself showed experimentally that neutrons are emitted in the process, and envisaged a divergent chain reaction with a vast release of energy. This, he realized, could result in an explosion of unprecedented power.

Rotblat moved to England in 1939 to work under James Chadwick in Liverpool, first on the university's new cyclotron particle accelerator, and then on the feasibility of the atomic bomb. In Poland he had married a student of literature, Tola Gryn, and he returned to Warsaw to fetch her. But she developed appendicitis, and Rotblat had to leave for England again on his own. Before she could follow him, Germany invaded Poland and war began. Despite all his efforts he was unable to get her out. She died during the war without his seeing her again.

In 1943, Rotblat followed Chadwick to Los Alamos National Laboratory in New Mexico, to work on the Manhattan Project that developed the atomic bomb. The project's morality disquieted him, but he feared that the Germans would develop the bomb first, and believed the Allies must be able to threaten retaliation. When intelligence showed that German progress was minimal, he resigned from the Manhattan Project on grounds of conscience and returned to Liverpool. After the war he took British citizenship, deciding not to return to communist Poland.

The bombing of Hiroshima and Nagasaki

in 1945 appalled Rotblat, and his life's mission began. He worked at first through the Atomic Scientists Association to educate the public about nuclear matters, and campaigned for the international control of nuclear energy. He switched his research to the medical applications of nuclear physics and joined the staff of St Bartholomew's Hospital Medical College at the University of London in 1949, becoming professor of physics in 1950. There he explored the use of linear accelerators for radiotherapy, and produced several landmark studies with Patricia Lindop on the effects of high-energy radiation on living tissue. But it was an American bomb test in 1954, which showered a Japanese fishing boat with radioactive fallout, that made Rotblat an international figure. He calculated that the bomb had been vastly more 'dirty' than the public had been told. His move to bring this matter into the open horrified government circles, which considered that all nuclear matters should be secret.

Around this time Rotblat met Bertrand Russell, who also was becoming increasingly concerned about the hydrogen bomb. In 1955 Rotblat was one of 11 prominent signatories of the Russell–Einstein Manifesto, a stark statement of the dangers of nuclear war. The manifesto led to the initial Pugwash Conference in 1957, in the village of Pugwash on the northern shore of Nova Scotia, at which scientists from across the world gathered to discuss how to avert a nuclear catastrophe. It was the first of more than 300 international conferences and workshops, in which participants speak as individuals whose remarks are unattributable. Throughout his life, Rotblat was a driving force in the Pugwash organization, becoming its secretary-general (1957–73), president (1988–97) and emeritus president.

Since its inception, Pugwash has been one of the foremost advocates of détente and disarmament in the nuclear age. It kept lines of communication open during the cold war and helped lay the foundation for important arms-control treaties. It provided the first links between Henry Kissinger and the North Vietnamese in the Vietnam war, and was an informal channel for officials and public figures in the Arab–Israeli, Korean and Kashmiri conflicts. In 1995, the Nobel Peace Prize was awarded jointly to Pugwash and to Joseph Rotblat.

But Rotblat's pacifist activities extended beyond Pugwash. He co-founded the UK Campaign for Nuclear Disarmament and was



the initiator of the Stockholm International Peace Research Institute. He participated in the Medical Exchange Programme between Britain and the Soviet Union, and was largely responsible for the comprehensive reports of the World Health Organization of 1984 and 1987 on the effects of nuclear war on health and health services. Shortly before his death, Rotblat had become increasingly concerned about developments in nuclear policy, particularly in the United States. He contacted leaders of other non-governmental organizations to initiate the Weapons of Mass Destruction Awareness Programme, launched in London in 2004 by himself and the former Soviet president Mikhail Gorbachev.

In addition to numerous papers on nuclear physics and radiation biology, Rotblat wrote, or co-wrote, more than 40 books on various aspects of the control of nuclear weapons and the prevention of war. Alongside the Nobel prize, he received the Bertrand Russell Society Award in 1983 and the Albert Einstein Peace Prize in 1992. Among British honours, he was appointed a Commander of the British Empire in 1965, and knighted in 1998.

Rotblat was a towering figure in the struggle for peace. He was brilliant, energetic, determined and eloquent: a man of utter integrity and great humanity, who committed his life to the pursuit of a saner, safer world. In his autobiography, Bertrand Russell said of Rotblat's work for disarmament: "If ever these evils are eradicated, his name should stand very high indeed among the heroes."

Sally Milne and Robert Hinde

Sally Milne and Robert Hinde are on the executive committee of Pugwash UK, 63A Great Russell Street, London WC1B 3BJ, UK. e-mail: pugwash@mac.com

M. PELLETIER/CORBIS SYGMA

BRIEF COMMUNICATIONS

Liquefaction of quicksand under stress

A person trapped in salt-lake quicksand is not in any danger of being sucked under completely.

People or animals caught in quicksand find it very hard to escape¹. Here we show that quicksand acts as a trap because it becomes unstable when it is forced to move — first it liquefies, and then it collapses. But a simple sinking test demonstrates that it is impossible for a human to be drawn into quicksand altogether.

The natural quicksand that we study here consists of fine sand, clay and salt water. Rheometrical tests (Fig. 1a, b) reveal its extreme sensitivity to very small variations in stress. At rest, its viscosity slowly increases with time — a behaviour characteristic of clays^{2,3}. This reflects the formation of a fragile colloidal gel that has a random, delicately balanced structure. At higher stress, a spectacular liquefaction of the material takes place: the steady-state viscosity changes by several orders of magnitude for a variation in stress of less than 1%. The higher the stress, the more liquid the quicksand becomes, so movement by a

trapped body causes it to sink in deeply.

Why is it that, once sunk in quicksand, it is so difficult to escape? Because the apparent viscosity of quicksand increases after the initial stress-induced liquefaction, unlike that of clay or sand alone^{3,4}. After liquefaction, the quicksand is seen to segregate into a water-rich phase and a sand-rich one. The apparent viscosity increase is therefore due to the formation of sand sediment, which has a very high volume fraction ($\phi \approx 0.8$) and viscosity. It is the difficulty of moving this densely packed, wet sand that leads to trapping. Water must be introduced into the compacted sand to liquefy it, which requires huge forces: to introduce water at a speed of 1 cm s^{-1} , say, a pressure of 10^6 pascals (Pa) is needed¹, assuming a typical sand-pore size of $10 \text{ }\mu\text{m}$. To pull out a foot at this speed, a force of some 10^4 newtons is required — about that needed to lift a medium-sized car.

By mixing sand and clay in salt water, a laboratory quicksand can be created with a structure that reproduces the behaviour of natural quicksand. It is just strong enough to support the weight of an adult person¹ at a very low volume fraction of sand ($\phi \approx 0.4$): the corresponding stress of about $5 \times 10^4 \text{ Pa}$ is similar to the measured elastic modulus of quicksand (Fig. 1c). This very loosely packed sand does not collapse under its own weight owing to the yield-stress of the colloidal clay gel. However, if the delicately balanced structure is perturbed, the gel will liquefy, rendering the packing of the sand unstable and leading to collapse⁵. Salt is an essential ingredient for the collapse in laboratory and natural quicksand — the latter originates from salt lakes whose salinity is close to that of the Dead Sea. The salt destabilizes the colloidal gels, causing the colloids to flocculate², which subsequently destroys the granular network.

We also simulated someone moving in quicksand to see whether — once partially submerged — the victim would

sink helplessly beneath the surface. A sinking test⁶ was used in which the speed at which an aluminium bead (radius $r = 2 \text{ mm}$) sinks into quicksand is measured. At rest, the bead remains on the surface, although it has a higher density (ρ) than the quicksand (2.7 g ml^{-1} compared with 2 g ml^{-1}). If the whole system is mechanically shaken to mimic movement in the quicksand, the results agree with the rheological findings (Fig. 1a, b). At small amplitudes (acceleration $a < 3.16 \text{ m s}^{-2}$), the bead stays afloat; however, liquefaction occurs at larger amplitudes and the resulting low viscosity causes the bead to fall to the bottom of the container (Fig. 1d). Liquefaction is so rapid in this case that sedimentation does not have time to occur.

Viscosity values differed for the rheology and sinking experiments as the initial states were different: in the sinking test, the sample had been allowed to age to enable it to support the bead. However, the critical acceleration does give roughly the same critical stress (exerted by the bead) for liquefaction as the rheology measurement of 1.3 Pa : $\Delta p r a / 3$ was about 1.5 Pa .

The most important conclusion from the sinking experiment is that it is impossible to sink beads with a density of 1 g ml^{-1} : they continue to 'float'. As this is typically the average density of humans and animals, any unfortunate victim should sink halfway into the quicksand, but could then take solace from the knowledge that there would be no risk of being sucked beneath the surface.

A. Khaldoun*, E. Eiser†, G. H. Wegdam*, Daniel Bonn*‡

*Van der Waals-Zeeman Institute and †HIMS, University of Amsterdam, 1018 XE Amsterdam, The Netherlands

e-mail: bonn@science.uva.nl

‡Laboratoire de Physique Statistique de l'Ecole Normale Supérieure, 75231 Paris Cedex 5, France

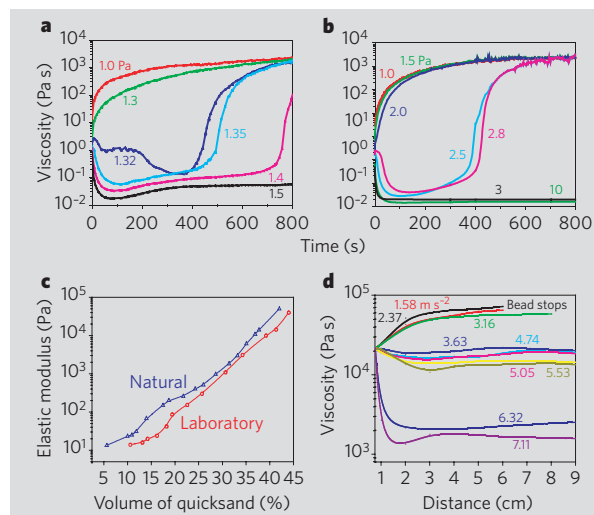


Figure 1 | Rheological and mechanical determination of quicksand properties. **a**, Liquefaction under shear of natural quicksand from a salt lake near Qom, Iran. Viscosity is plotted against time for quicksand (water content, 50% by weight (wt%); grain size, 50–200 μm ; clays, about 7 wt%, mostly montmorillonites; salinity, 0.1 M) for the imposed stress levels indicated in the figure. **b**, As **a**, but with laboratory quicksand (90 wt% sand, 10 wt% bentonite in salt water; total water, 50 wt%). Salinity higher than 0.02 M is necessary for collapse, which is visible as a viscosity increase after liquefaction. **c**, Shear elastic modulus, G' , of natural and of laboratory quicksand for different volume fractions of water, measured with a rheometer (frequency, 1 Hz; deformation, 0.1%). **d**, Sinking experiment, showing viscosity as a function of depth of sinking in a quicksand column (50% water) for different amplitudes of shaking. For comparison with results in **a**, **b**, we converted the falling speed into an effective viscosity by using Stokes law⁷.

1. Lambe, T. W. & Whitman, R. V. *Soil Mechanics* (Wiley, New York, 1969).
2. Van Olphen, H. J. *An Introduction to Clay Colloid Chemistry* (Wiley, New York, 1977).
3. Coussot, P., Nguyen, Q. D., Huynh, H. T. & Bonn, D. *Phys. Rev. Lett.* **88**, 175501 (2002).
4. Huang, N. et al. *Phys. Rev. Lett.* **94**, 028301 (2005).
5. Lohse, D., Rauhé, R., Bergmann, R. & van der Meer, D. *Nature* **432**, 689–690 (2004).
6. Ferroir, T., Huynh, H. T., Chateau, X. & Coussot, P. *Phys. Fluids* **16**, 594–601 (2004).
7. Batchelor, G. K. *An Introduction to Fluid Dynamics* (Cambridge Univ. Press, Cambridge, 1967).

Competing financial interests: declared none.
doi:10.1038/437635a

ROBOTICS

Self-replication from random parts

Autonomously self-replicating machines have long caught the imagination^{1–3} but have yet to acquire the sophistication of biological systems, which assemble structures from disordered building blocks. Here we describe the autonomous self-replication of a reconfigurable string of parts from randomly positioned input components. Such components, if suitably miniaturized and mass-produced, could constitute self-fabricating systems whose assembly is brought about by the parts themselves.

A key feature of biological replication is a template molecule's ability to make copies of itself (as in the case of DNA) by selecting the appropriate building blocks (nucleotides) from parts that are randomly and continuously distributed in its environment; the system also has a built-in ability to correct errors made during copying⁴. The efficiency of this two-step process enables biological systems to generate exponential numbers of accurate copies of themselves as a function of time. To create these properties in an artificial system, a machine needs to be capable of autonomous acquisition of randomly distributed building blocks and of carrying out error correction during the copying process.

A scheme for the autonomous self-replication of a simple 2-bit mechanical string was outlined almost half a century ago³. Replication of more complex systems using structured inputs has since been achieved^{5–7}, including a self-reproducing machine that relies on a well-ordered supply of its building blocks⁸.

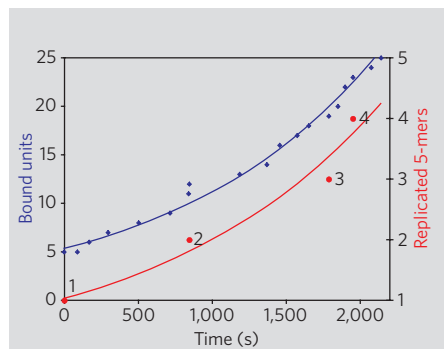


Figure 2 | Replication kinetics for bound building blocks (blue) and completed strings (red) as a function of time. The kinetics of these processes are exponential until they become limited by the supply of component parts. Strings are numbered as in Fig. 1.

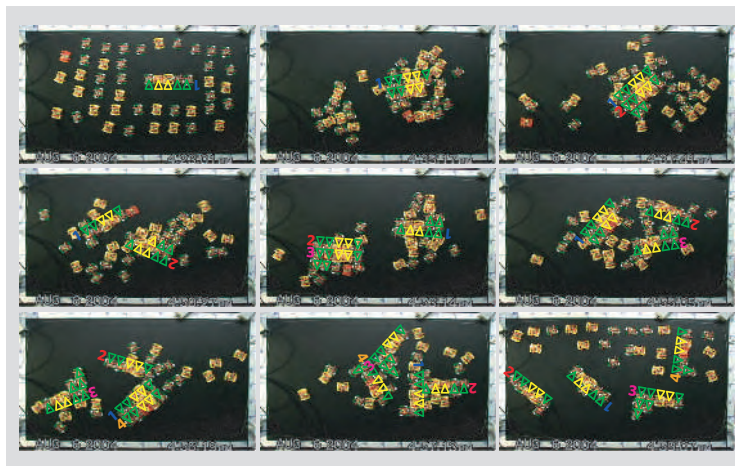


Figure 1 | Self-replication of a 5-bit string. Frames 1–9 (from left to right in each row): time sequence of photographs showing the autonomous replication of a 5-bit string of electromechanical units, starting from a single initial input string (number 1: green, green, yellow, yellow, green); frame 3: multiple replicants (numbered 1,2) assembling on a single substrate. Addition of building blocks is purely sequential along the string, as governed by a rule running in each block's state machine; frame 9: four independent strings (numbered 1–4) result from the action of templating and division. For movie, see supplementary information.

Our focus is on the autonomous replication of complex systems from random inputs. The complexity of a given structure may be defined by the bit length describing the configuration of parts — in this case, a 5-bit string. If the error per addition (arising from random input) of each new building block in the copied string is ϵ , then the yield for replicating an n -bit string is $(1 - \epsilon)^n$, which becomes exponentially small for complex (large n) systems ($\epsilon = 0.5$, $n = 5$; the yield is about 3% in the case described here).

For complex structures to be copied accurately from random inputs, some mechanism for error correction is required⁹; here, error correction is defined as a process in which a linear increase in resource leads to an exponential decrease in error rate. In DNA replication, for example, the polymerase enzymes responsible for copying may also check each recruited nucleotide base for correct complementary base-pairing with the DNA template strand: if the incoming base does not fit, it is removed by the enzyme's exonuclease domain.

To implement error-correcting replication in an artificial system, we constructed a set of programmable electromechanical components that are run as a 7-state, finite-state machine; the components can be reversibly latched and unlatched in response to nearest-neighbour communications¹⁰ (see supplementary information). These parts interact by floating on a two-dimensional air table on which motion is random.

Figure 1 shows a series of frame shots that start with a single-seed string (coloured in a green, green, yellow, yellow, green sequence).

Self-replication of this sequence occurs as a result of a random part latching on to the seed string, then the part is queried for self-similarity and proper position in the growing replicant, and subsequently it is either permanently latched or released according to an embedded rule (for movie, see supplementary information). We note that each part needs only to run a local, compact (7-state) state machine and does not, in itself, store an entire copy of the structure. The kinetics of these processes are exponential until they become limited by the supply of parts (Fig. 2).

In addition to replication, several other algorithms, including one-dimensional (line) and two-dimensional (checkerboard) pattern formation from internal rules, as well as a reconfigurable pattern formation, were run using the same system (see supplementary information). Controls run without error correction, however, show large (random) error rates in the final assemblages.

Given the compact requirements of internal-state machines, coupled with recent advances in microelectromechanical systems, it is possible that components such as those we describe here could eventually be miniaturized. They might be used to create a general system that is capable of self-replicating or of being programmed to self-fabricate into complex structures that run with exponential kinetics.

Saul Griffith, Dan Goldwater, Joseph M. Jacobson

Center for Bits and Atoms, MIT Media Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
e-mail: jacobson@media.mit.edu

1. Von Neumann, J. *Theory of Self-Reproducing Automata* (Univ. Illinois Press, Champaign, Illinois 1966).
2. Freitas, R. A. Jr & Merkle, R. C. *Kinematic Self-Replicating Machines* (Landes Bioscience, Georgetown, Texas, 2004).
3. Penrose, L. S. & Penrose, R. *Nature* **179**, 1183 (1957).
4. Alberts, B. et al. (eds) in *Molecular Biology of the Cell* 4th edn, 241 and 983 (Garland Science, New York, 2001).
5. Jacobson, H. *Am. Sci.* **46**, 255–284 (1958).
6. Penrose, L. S. & Penrose, S. V. *Automatic Mechanical Self-Replication* (2) (Cresswell Film Unit, Galton Laboratory, University College London, December 1961).
7. Suthakorn, J., Cushing, A. B. & Chirikjian, G. S. *Proc. 2003 IEEE/ASME Int. Conf. Adv. Intel. Mechatronic (AIM)* (Kobe, Japan, 2003).
8. Zykov, V., Mytilinaios, E., Adams, B. & Lipson, H. *Nature* **435**, 163–164 (2005).
9. Banks, E. R. *Information Processing and Transmission in Cellular Automata*. Thesis, Massachusetts Institute of Technology (1971).
10. Griffith, S. *Growing Machines*. Thesis, Massachusetts Institute of Technology (2004).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
doi:10.1038/437636a

A perspective on surfaces and interfaces

David L. Allara¹

The importance of surfaces and interfaces cannot be overstated, with their reach extending from the hardware of the digital age to the processes of life. The past half-century has seen the development of a full and varied toolkit for characterizing them. This toolkit is now serving a growing interdisciplinary community and is providing a powerful platform for scientific research and manufacturing technology.

Surfaces and interfaces are ubiquitous. They are found in systems as simple as a piece of metal in a vacuum, and as complex as biological cells and living organisms. They define a boundary with the surrounding environment and influence interactions with that environment, and so it is no surprise that interfaces have been appreciated historically — just think of how corrosion, tarnishing and friction have plagued the hardware of civilization. But a cursory scan of the scientific and technological literature shows that the direct study of real surfaces is a fairly recent phenomenon. It is only over the past decade or so that the subject of interfaces has moved to the forefront of an increasing number of fascinating fundamental scientific inquiries. The ability to precisely engineer interfaces is playing an increasingly dominant part in the development of new technologies relevant to all aspects of our lives, from energy production to biomedical implants.

I think it is useful to look back and consider how we arrived at this point, not least because it serves as an interesting example of how science often evolves slowly during years of patient study, followed by a sudden explosion in the number of new insights and new applications. In the case of interface science, it is not that many years ago when there still were no tools available to us for directly interrogating the tiny amounts of matter present in surfaces and interfaces. The mass balance, which had been such a powerful instrument for early chemists, was incapable of measuring the mass of a surface layer. But today, it is almost taken for granted that we can directly image and even control single atoms and molecules on a surface and create useful new structures. These endeavours are supported by continuously evolving theories, which are in turn bolstered by the dizzying increase in the power of computers so that simulations now routinely help unravel the details of interface phenomena, such as the behaviour of fluids in confined spaces while flowing across chemically structured surfaces. When I consider these achievements and my journey with a large number of colleagues down the path of surface and interface research over the past four decades, I am truly struck by the confluence of what at one time were considered wholly different streams of science. The articles that follow in this issue give an excellent demonstration of how the merging of different scientific streams has given us a commanding toolbox, which makes it possible to advance the frontiers of interface science and technology in fields as diverse as electronics, cell biology and sensor development.

In my mind, the need to understand in detail surfaces and interfaces and to control them really heightened in the first half of the past century. At that time, machinery — particularly automobiles — became an increasingly important factor in our economy so that developing methods for controlling phenomena such as friction, lubrication, adhesion, wetting, corrosion and surface oxidation provided opportunities for enormous economic gains. Although incisive

experimental tools for probing surfaces were developed only later, scientists and engineers nevertheless had useful chemical and physical concepts on hand to guide their thinking. In fact, a number of important surface and interface-related phenomena were uncovered during this period, including the mechanism of the photo-electric effect and the invention of the transistor. Reasonably accurate concepts were also developed for interfacial phenomena involving soft matter, such as the self-organization of a monolayer of amphiphilic (surfactant) molecules at metal surfaces as relates to lubrication, wetting and adhesion^{1,2}.

New developments

But the true birth of surface and interface science, where molecular and atomic details of a surface are imaged and manipulated directly, occurred only in the second half of the past century. As has been nicely outlined in a historical perspective by Duke³, the birth and subsequent evolution of surface science were driven by technological innovations. Only when ultra-high vacuum systems became available in the 1960s was it possible to create and maintain well-defined surfaces. Still, the direct and quantitative determination of the atomic composition and structure of clean metal surfaces under vacuum conditions had to await the development of electron and ion spectroscopies, which occurred in electrical engineering and physics labs during the 1960s to 1980s. These efforts to uncover the atomic details of surfaces coincided with investigations of molecular self-assembly at surfaces, which started in the early 1980s and was enabled in good part by the emergence of photon-based surface characterization tools, such as monolayer sensitive infrared, optical and photoelectron spectroscopies.

During this time, I was working on surface molecular assembly at Bell Laboratories, surrounded by many of the vacuum-surface science pioneers of the time. But our direct interaction was really surprisingly small, with the experimental hardware and concepts needed to drive the vacuum side of surface science firmly rooted in physics and the molecular side in chemistry⁴. What overlap there was proved, of course, to be extremely stimulating. Overall, however, vacuum-surface science — or 'hard' surface science because of its focus on bare single crystals of metal — was evolving along its own course, while molecular surface assembly headed off in another direction to become the 'soft' surface science concerned with the behaviour of molecules such as surfactants and even polymers at interfaces⁵.

Developments in the hard surface science have given us ever faster computers and communication technologies. Evolution of the soft interface science has opened avenues for studying biological interfaces and, in the late 1980s, kick-started 'soft lithography' as a simple and versatile lab-bench method for chemical patterning of surfaces down to submicrometre dimensions. Intriguingly, processing capabilities of hard surface science have been combined with soft lithography to con-

¹Department of Polymer Science and Chemistry, 206 Chemistry Building, University Park, Pennsylvania State University, Pennsylvani, USA.

trol the wetting behaviour of fluids under confinement and on chemically patterned surfaces. These so-called microfluidic systems, which first emerged in the early 1990s, are now mainly created by simple soft lithography methods and have been applied to a wide range of analytical and sensing systems. Yet another stream of scientific development appeared in the mid to later 1980s, when chemists began to learn to control the precipitation of simple inorganic compounds from solution to create uniform, nanometre-sized crystallites. Such crystallites or colloidal particles with nanometre-scale dimensions have been produced and used throughout history. For example, gold colloids were used for decorating pottery or staining glass, and silver colloids formed the basis of photographic film. But transforming the production of particles from a highly empirical art to a rational, adjustable method was only possible when the principles developed to explain wetting and surface molecular assembly were used to explore the factors that control the sizes, shapes and properties of colloidal nanoparticles. This has made it possible to achieve impressive control over nanoparticle formation, to the extent that we can now produce nanoparticles with narrow size distributions, alter the shape of particles by selective growth of appropriate crystal faces and tune particle properties such as their optical response.

The contemporary toolkit

While each of these several fields was evolving separately, the ability to routinely observe and even manipulate individual objects at the nanometre scale remained tantalizingly out of reach. Although electron microscopy had been around for decades, it could not be used for many samples and problems. The advent of scanning probe microscopies (SPM) filled this gap and has over the course of the 1990s and into the present, revolutionized surface and interface science (and, incidentally, seems to have prompted the arrival of 'nanotechnology' as the label of choice for every study looking at something 'small'). The first SPM studies looked at 'hard' surfaces that are of interest for microelectronics and heterogeneous catalysis, revealing the atomic structure of single crystals held in vacuum. Studies then also shifted to 'soft' systems to explore self-assembled molecules at surfaces. Nowadays, even surfaces immersed in liquid can be imaged using SPM. This capability, along with the emergence of non-linear laser spectroscopies that are sensitive to wet interfaces, has started to provide incisive access to biological interface problems.

Another crucial aspect of the development of surface and interface science is the emergence of computational tools that make it possible to use evolving theory, ranging from quantum to statistical mechanics, to tackle simulations and analyses of enormously complex interfacial behaviour. In a field that has mainly emerged from empirical observations and experiment, it is now not unusual to see that well-executed theory and simulation can be accepted as more useful than experiment (which is often prohibitively costly and difficult for complex systems and phenomena).

What seems amazing to me is that hard and soft surface and interface science have delivered a powerful range of common experimental and theoretical tools that are proving useful in areas as diverse as microelectronics and biocompatibility. Groups of scientists from diverse backgrounds have access to these tools for very different

investigations and can think in a common way about diverse surface and interface phenomena and applications. The following articles illustrate some of the tools, concepts and knowledge that are now readily available to those who study interfaces and engineer interfacial phenomena and structures for practical applications.

Chandler (p. 640) gives an example of how well-constructed theory can help to identify the basic structural and energetic factors that control the behaviour of water at hydrophobic surfaces and thereby develop a powerful yet simple understanding of an often complex phenomenon that affects a wide range of systems and processes. This approach to understanding the interfacial behaviour of liquids will accelerate progress in developing applications for microfluidics and bio-membranes, and bring the understanding of fundamental phenomena such as wetting to new levels. Considering engineered device structures, Atencia and Beebe (p. 648) show that the ability to build micrometre-to-nanometre-scale channel structures provides a means to exploit the fundamental principles of fluid behaviour in confined geometries. The result is a wide range of useful microfluidic devices that harness interface effects. Moving towards softer, biologically relevant structures, Tanaka and Sackmann (p. 656) detail strategies for constructing and using improved model cell membranes. These essentially self-organized molecular layers, tethered to surfaces, are coupled with analytical probes to study fundamental processes occurring in or on biological membranes and developed for sensor applications. These last two papers point to a distinct change in the ability to use fundamental knowledge for bio-engineering. Yin and Alivisatos (p. 664) review the rapid progress that has been made in forming nanometre-sized crystallites of inorganic materials with excellent control over their sizes and shapes. They show that fundamental kinetic and thermodynamic principles, along with judicious use of molecular adsorption at the crystallite surface, allow us to select growth pathways to achieve the controlled formation of unprecedentedly complex inorganic nanostructures. Finally, Barth, Costantini and Kern (p. 671) give us a look at how the traditional area of vacuum-surface science with single crystal surfaces has evolved into a highly sophisticated art. Again, kinetic and thermodynamic principles are used to precisely control the formation of complex ordered surface structures that might find use in the information industry. Overall these articles underscore the importance of the confluence of surface and interface research in recent years and point the way to future developments and applications. ■

1. Bowden, F. P. & Tabor, D. in *The Friction and Lubrication of Solids* (Oxford Univ., Oxford, 2001).
2. Baier, R. E., Shafrin, E. G. & Zisman, W. A. Adhesion, mechanisms that assist or impede it. *Science* **162**, 1360–1368 (1968).
3. Duke, C. B. The birth and evolution of surface science: Child of the union of science and technology. *Proc. Natl Acad. Sci. USA* **100**, 3858–3864 (2003).
4. Ulman, A. Formation and structure of self-assembled monolayers. *Chem. Rev.* **96**, 1533–1554 (1996).
5. Love, J. C., Estroff, L. A., Kriebel, J. K., Nuzzo, R. G. & Whitesides, G. M. *Chem. Rev.* **105**, 1103–1169 (2005).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence and requests for materials should be addressed to D.L.A. (dla3@psu.edu).

Interfaces and the driving force of hydrophobic assembly

David Chandler¹

The hydrophobic effect — the tendency for oil and water to segregate — is important in diverse phenomena, from the cleaning of laundry, to the creation of micro-emulsions to make new materials, to the assembly of proteins into functional complexes. This effect is multifaceted depending on whether hydrophobic molecules are individually hydrated or driven to assemble into larger structures. Despite the basic principles underlying the hydrophobic effect being qualitatively well understood, only recently have theoretical developments begun to explain and quantify many features of this ubiquitous phenomenon.

Most general chemistry students are taught how detergent removes grease. An essential aspect of this process depends on the amphiphilic nature of the detergent molecules. Detergent molecules contain polar or charged components that happily interact with water (hydrophilic), and apolar components that by themselves do not easily dissolve in water (hydrophobic). Although these contrasting components would normally separate into a water-rich phase and an oil-rich phase, placing them next to each other in one molecule suppresses this macroscopic separation. Instead, the detergent molecules aggregate into mesoscopic fluid structures, such as micelles, with oily interiors and watery exteriors. Grease dissolves in aqueous detergent solutions by dispersing into the interiors of these aggregates. Similar effects influence more complicated assemblies¹, including biological structures where the separation of hydrophobic and hydrophilic components is a common feature².

In all these systems, an interaction mediated by water — the hydrophobic interaction — seems to cause clustering of hydrophobic units. This was noted by Walter Kauzmann in his influential 1959 paper³, and the idea is now widely accepted^{4,5}. But quantifying it in simple terms has been difficult because hydrophobicity is a multifaceted phenomenon that manifests different characteristics depending on whether small molecular units or large clusters are involved, or a combination of both. Here, I review these different regimes and theory to deal with them.

Contrasting length scales

The segregation of hydrophobic and hydrophilic phases or components results in a molecular interface that extends over distances (or 'lengths') that are large compared with the distances over which molecules affect one another in a homogeneous liquid. These interfaces distinguish mesoscopic structures, such as micelles, bilayers or microemulsions, from essentially nondescript intermolecular arrangements. For example, a few hydrated methane molecules and the alkyl groups of small alcohol molecules mixed with water show insignificant tendencies to cluster^{6,7}. Although the relative positions of small hydrophobic molecules in water are correlated, these correlation effects are modest and similar to what is seen for most small molecules in most homogeneous liquids. Such correlations in relative position are not a result of clustering, but arise from the effects of molecular size and shape^{8–11}.

The distances over which molecules influence one another in a homogeneous liquid typically approximate the girth of a molecule: about 0.3–0.5 nm for water or a small alkane. On this length scale, that is, the bulk liquid correlation length, molecular reorganization occurs readily because it involves only a modest thermodynamic cost. But reorganization to form an interface involves a significant cost. And if the interface is to remain, this cost or free energy of formation needs to be compensated for by forces that favour separation of the system into different phases. Because the interface cost grows linearly with surface area, whereas the compensating forces grow linearly with volume, a cluster can be stable or metastable only if it exceeds some critical size. This argument is familiar in the context of nucleation theory, which notes that a super-saturated solution will phase separate only after the formation of a critical nucleus¹². To be long-lived, hydrophobic clusters must extend over a minimum length (of 1 nm or more; see below).

Because bulk driving forces and interfacial costs compete, a consideration of only one of these quantities is generally insufficient to explain hydrophobic effects. And although efforts to infer hydrophobic interactions from molecular-scale surface areas alone are common^{13,14}, experimental observations illustrate the limitations of this approach. For example, the free energy required to transfer small hydrophobic molecules from oil to water differs by a factor of three compared with that inferred from interfacial cost considerations¹⁵. Moreover, while interfacial costs decrease with increasing temperature, hydrophobic forces show the opposite trend³.

Because hydrophobic interactions increase in strength with increasing temperature, they are often viewed as entropic; that is, hydrophobic units induce some order in the surrounding water. This idea is correct in as much as small hydrophobic units reduce the volume of configuration space available for hydrogen bonding. But the extreme view that pictures hydrophobic solvation in terms of rigid clathrate structures, like those surrounding hydrophobic particles in gas hydrates, is clearly incorrect: intermolecular correlations in liquid matter are insufficiently strong to be consistent with this crystalline picture. And while remnants of clathrate structure persist in the liquid near a small hydrophobic particle¹⁶, a surrounding clathrate structure is geometrically implausible in the case of extended hydrophobic surfaces.

¹Department of Chemistry, University of California, Berkeley, California 94720, USA.

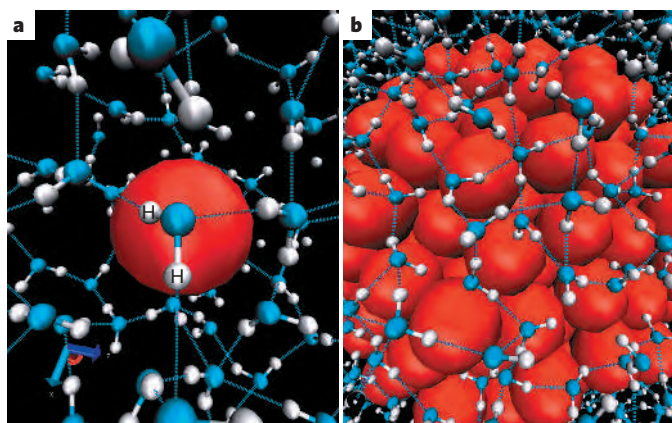


Figure 1 | Configurations of liquid water molecules near hydrophobic cavities in molecular-dynamics simulations. The blue and white particles represent the oxygen (O) and hydrogen (H) atoms, respectively, of the water molecules. The dashed lines indicate hydrogen bonds (that is, O-H...O within 35° of being linear and O-to-O bonds of no more than 0.35 nm in length). The space-filling size of the hydrophobic (red) particle in **a** is similar to that of a methane molecule. The hydrophobic cluster in **b** contains 135 methane-like particles that are hexagonally close-packed to form a roughly spherical unit of radius larger than 1 nm. In both cases, the water molecules shown are those that are within 0.8 nm of at least one methane-like particle. For the single cavity pictured in **a**, each water molecule can readily participate in four hydrogen bonds. (Owing to thermal motions, hydrogen bonding in liquid water is disordered.) Water molecules in **a** are typical of the bulk liquid where most molecules participate in four hydrogen bonds. The water molecules shown in **b**, however, are not typical of the bulk. Here, the cluster is sufficiently large that hydrogen bonds cannot simply go around the hydrophobic region. In this case, water molecules near the hydrophobic cluster have typically three or fewer hydrogen bonds.

Hydration of small and large cavities

Figure 1 illustrates this point by juxtaposing pictures of the hydration of small and large hydrophobic solutes. Most molecules dissolved in water have complicated shapes. Nevertheless, idealized spherical apolar particles or nearly spherical clusters, as pictured here, capture the most important physical features responsible for hydrophobic effects: acting like cavities in the water, these solutes exclude water molecules from the volumes they occupy, and they present regions of space where hydrogen bonding cannot occur.

Hydrophobic molecules interact with water in a variety of ways in addition to excluding volume. They exert weak attractive forces on water molecules by means of van der Waals interactions. They also exert strong attractive forces via hydrophilic components, such as the hydroxyl group on an alcohol. Although van der Waals interactions are too weak to affect the existence of interfaces in water, they do affect the position of an interface. Similarly, whereas hydrophilic parts of amphiphilic molecules are not directly responsible for hydrophobic assemblies, they do affect the arrangement of these assemblies relative to interfaces and other structures. We will look at both these effects, but first consider the most important physical features of hydrophobic solutes, all of which are found in the analysis of how cavities are solvated in water.

The small-solute case depicted in Fig. 1a illustrates the solvation of the cavity associated with a molecule such as methane. Namely, it excludes the centres of water molecules from a spherical volume less than 0.5 nm across. This volume is small enough that its presence in water requires no breaking of hydrogen bonds. Water molecules can adopt orientations that allow hydrogen-bonding patterns to go around the solute, and the extent to which bonds are broken at any instant is similar to that in the pure liquid. The situation is different in the large-solute case illustrated in Fig. 1b. Here, the solute surface extends with low curvature over areas larger than 1 nm², making it impossible for adjacent water molecules to maintain a complete hydrogen-bonding

network with the surrounding liquid. A fraction of the hydrogen-bonding possibilities are thus lost near an extended hydrophobic surface. To minimize the loss, on average, less than one hydrogen bond per water molecule is sacrificed compared with that in the bulk liquid. As a result, water tends to move away from the large solute and forms an interface around it akin to that between liquid and vapour.

This idea — that hydrogen bonding is maintained near a small hydrophobic region and not maintained near a large hydrophobic region — was expressed more than 30 years ago by Frank Stillinger¹⁷. It provides the physical basis for understanding hydrophobic effects.

Thermodynamics

Thermodynamic costs indicate whether processes are likely to occur and are conveniently quantified in terms of a free energy ΔG . In the context of solvating a molecule, ΔG is the reversible work for the solvent to reorganize and solvate the solute. The probability of solvation happening is proportional to $\exp(-\Delta G/k_B T)$, where T is temperature and k_B is Boltzmann's constant. This principle of statistical mechanics¹¹, which relates reversible work to probability, allows ΔG to be determined by measuring equilibrium constants, such as the concentration of the solute in water relative to that in some other environment. The free energy ΔG can also be computed using microscopic theory (see Box 1).

The free energy has two primary components: $\Delta G = \Delta H - T\Delta S$, where ΔH and ΔS are the enthalpic and entropic changes incurred during solvation. The enthalpic part is a measure of the average potential energy of interaction between molecules, and the entropic part is a measure of the order or intermolecular correlations¹⁸. The free energy of a process involving significant changes in the number of molecular interactions, such as the breaking of hydrogen bonds to form a liquid–vapour interface, will be dominated by its enthalpic component. In such cases, $\Delta G/T$ will decrease with increasing temperature.

A process that requires specific spatial organization of hydrogen-bonding patterns will have an important entropic component. At room temperature, for instance, the entropic cost of hydrating small hydrophobic species is dominant, as manifested by ΔG increasing with increasing temperature. With a sufficient increase in temperature, however, the extent of hydrogen bonding between water molecules diminishes, and maintaining hydrogen bonds becomes less important. In fact, whereas the entropy change associated with the hydration of a

Box 1 | Calculating solvation energies

Dissolving a substance in a solvent can be regarded as transforming a system from state 1 (pure solvent) to state 2 (solvent plus solute). This process is associated with a change in free energy, $\Delta G = G_2 - G_1$, which in our example is the solvation free energy. Macroscopic properties such as ΔG can be determined from the molecular properties and molecular interactions of the system (captured through so-called partition functions Z) using statistical thermodynamics¹¹:

$$\Delta G = G_2 - G_1 = -k_B T \ln(Z_2/Z_1) \quad (1)$$

$$= -k_B T \ln(\langle \exp(-\Delta E/k_B T) \rangle_1) \quad (2)$$

$$\approx \langle \Delta E \rangle_1 \quad (3)$$

Here, ΔE denotes the difference in microstate energy between states 2 and 1, and $\langle \dots \rangle_1$ denotes the equilibrium ensemble average (that is, the Boltzmann-weighted average) over the microstates of state 1. The third approximate equality is valid when $\Delta E/k_B T$ is predominantly small in the ensemble of microstates. The averages in equations (2) and (3) can be computed in a number of ways⁵⁹ to obtain ΔG .

A 'Boltzmann weight' of a microstate is the exponential of the energy in units of $-k_B T$. Thus, this thermal energy $k_B T$ is the energy scale of statistical thermodynamics against which energies or free energies are described as 'small' or 'large'. Microstates with small energy differences have similar probabilities and the more such states near a given energy, the larger the entropy. A partition function for a specific macroscopically controlled state is the Boltzmann-weighted sum over all microstates consistent with that control. The entropic contribution to free energy comes from the number of terms in that sum available at a given energy.

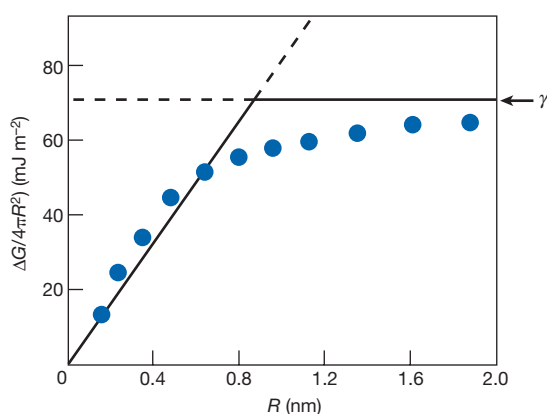


Figure 2 | Solvation free energy, ΔG , for a spherical cavity in water as a function of the cavity size. The results are for ambient conditions (room temperature and 1 atm pressure). The circles show the results of detailed microscopic calculations²⁵. The liquid–vapour surface tension is shown by γ . The solid lines show the approximate scaling behaviour of $\Delta G/4\pi R^2$ for small R , and the asymptotic behaviour for large R . This approach can be used to infer the typical length characterizing the crossover behaviour, but not the quantitative behaviour of ΔG in the crossover regime.

Box 2 | Free energy for small cavities in water

For a cavity with volume v (which need not be spherical), the difference in microstate energy ΔE is infinite whenever a solvent particle is in the cavity, and zero otherwise. This implies that the solvation energy ΔG_v of the cavity depends on $P_v(N)$, the probability of finding in a volume v of pure solvent, N solvent molecules. According to equation (2) from Box 1, $\Delta G_v = -k_B T \ln P_v(0)$. For small volumes, the probability $P_v(N)$ is almost exactly gaussian⁶⁰, and ΔG_v can therefore be expressed analytically in terms of the cavity volume v , the mean number of molecules that occupy that volume in the pure liquid $\langle N \rangle_v (= v\rho$, where ρ is solvent density), and the mean-square fluctuation in that number of molecules χ_v . In particular we find that⁵⁰:

$$\Delta G \approx k_B T \rho^2 v^2 / 2\chi_v + k_B T (\ln 2\pi\chi_v) / 2 \quad (4)$$

with

$$\chi_v = \langle (\delta N)^2 \rangle_v = \rho v + \rho^2 \int_v \int_v d\mathbf{r} d\mathbf{r}' [g(|\mathbf{r} - \mathbf{r}'|) - 1] \quad (5)$$

The integrand for integrals over the cavity volumes in the second equality in equation (5) is the pair correlation function for the pure liquid. Given that the radial distribution function, $g(r)$, is unity beyond the correlation length of the liquid, it follows that χ_v is roughly proportional to v . This behaviour together with equation (4) explains why ΔG for small hydrophobic solutes is approximately linear in solute volume.

The temperature dependence obtained from equation (4) approximates the temperature dependence of ΔG for small hydrophobic molecules in water⁵⁰. For example, consider a cavity of diameter 0.35 nm, as would be appropriate for methane. The entropy of solvation, $\Delta S = -\partial \Delta G / \partial T$, calculated from equation (4) at room temperature, is 20 cal mol⁻¹ K⁻¹. The heat capacity of solvation, $\Delta C_p = -T \partial^2 \Delta G / \partial T^2$, computed from equation (4) for that same cavity in water, is about 50 cal mol⁻¹ K⁻¹. These numbers agree well with those measured for methane in water at room temperature, about 17 cal mol⁻¹ K⁻¹ and 50 cal mol⁻¹ K⁻¹ for the solvation entropy⁶¹ and heat capacity⁵¹, respectively.

Equations (4) and (5), due to Pratt and his coworkers^{50,60}, can be applied easily to cavities and yield accurate solvation free energies for hydrophobic molecules of arbitrary shape, provided the molecules are not large. The older theory⁸⁻¹⁰ of hydrophobic solvation and interaction has similar generality, a similar range of applicability, and a similar, although less transparent, basis in gaussian statistics⁶². When this approach fails, the failure is due to the formation of interfaces, in which case a more general theory is applicable²⁰. These theories do not require the extent of microscopic detail that is considered important in some approaches^{63,64}. Once a cavity is specified, whether spherical or otherwise, all quantities entering into formulae for solvation free energies and solute distribution functions are experimentally measurable.

small alkane at room temperature is negative (and reasonably large in magnitude), it becomes positive near the boiling temperature of water¹⁹.

Interfaces and size scaling of ΔG

At ambient conditions (room temperature and 1 atm pressure), liquid water lies close to phase coexistence with its vapour; that is, the free-energy difference between water in its liquid and vapour phases is small compared with the thermal energy available for molecules to move from the liquid to the vapour phase. This condition ensures that large cavities in water are accompanied by an interface like that between liquid and vapour, as suggested by Stillinger¹⁷, and confirmed by theoretical analysis^{20,21} and simulation²²⁻²⁶. The cost to hydrate the large spherical cavity of radius R is thus $\Delta G \approx 4\pi R^2 \gamma + (4/3)\pi R^3 p \approx 4\pi R^2 \gamma$, where γ refers to the liquid–vapour surface tension, and p to the pressure, both at the temperature considered. The pressure–volume term $(4/3)\pi R^3 p$ would be important for macroscopic cavities, but is negligible at standard pressures provided R is less than several nanometres.

Changing the thermodynamic state, so as to move away from liquid–vapour coexistence, reduces the tendency to form a liquid–vapour-like interface near large cavities. Furthermore, the demand for interface formation — to minimize the number of broken hydrogen bonds — is diminished at high pressures, high temperatures, or both because these conditions lead to high concentrations of broken hydrogen bonds, even in the absence of interfaces. At standard conditions, however, a large hydrophobic solute does induce the surrounding water to form an interface, so that its solvation free energy contains a component that is proportional to surface area.

In contrast, the hydration of a small solute does not lead to broken hydrogen bonds, but involves a re-ordering of hydrogen bonds. This re-ordering persists into the surrounding liquid for a distance of about one correlation length (see also Box 2). Because correlations in water extend over the girth of a typical small molecule, adding up the points over which a small solute will affect correlations in water gives a solvation free energy that scales more like the volume than like the surface area of the solute. The ΔG needed to hydrate a small sphere will thus scale more accurately as R^3 than as R^2 (Box 2).

Figure 2 illustrates how solvation free energy (normalized to surface area) changes with solute size, and how the different trends for small and large hydrophobic solutes change once a radius of about 1 nm is reached. For smaller solutes, the solvation free energy grows linearly with solute volume; for larger solutes, it grows linearly with surface area. The crossover behaviour arises because only larger solutes induce adjacent water to form an interface. However, this crossover phenomenon is not a phase transition where collective motion on a macroscopic scale results in sudden or singular changes. The crossover pertinent to hydrophobicity is collective, but it occurs on a microscopic scale where nothing so precipitous as a true transition can occur.

Hydrophobic molecules: wet or dry?

The behaviour of water near hydrophobic solutes can be probed in detail by microscopic calculations, which average behaviour over microstates of the solvent in the presence (or absence) of the solute (see also Box 1). For example, Fig. 3 shows the density of solvent, relative to that of the bulk solvent, near a series of hydrophobic solute models of different sizes. For the cavity with the smallest excluded volume radius, $R = 0.4$ nm, the water density immediately adjacent to the solute surface is increased by a factor of two. In this case, the solute is said to be ‘wet’. The significant increase in water density adjacent to this cavity results from the liquid responding elastically to the cavity: water molecules are localized so as to maintain hydrogen bonding. In contrast, the larger cavities with $R = 1, 10$ and 100 nm are ‘dewetted’ or ‘dry’ because the large cavity has caused the source of elasticity — the hydrogen bond network — to break, so that the liquid moves away. The similarity between the interfaces formed here

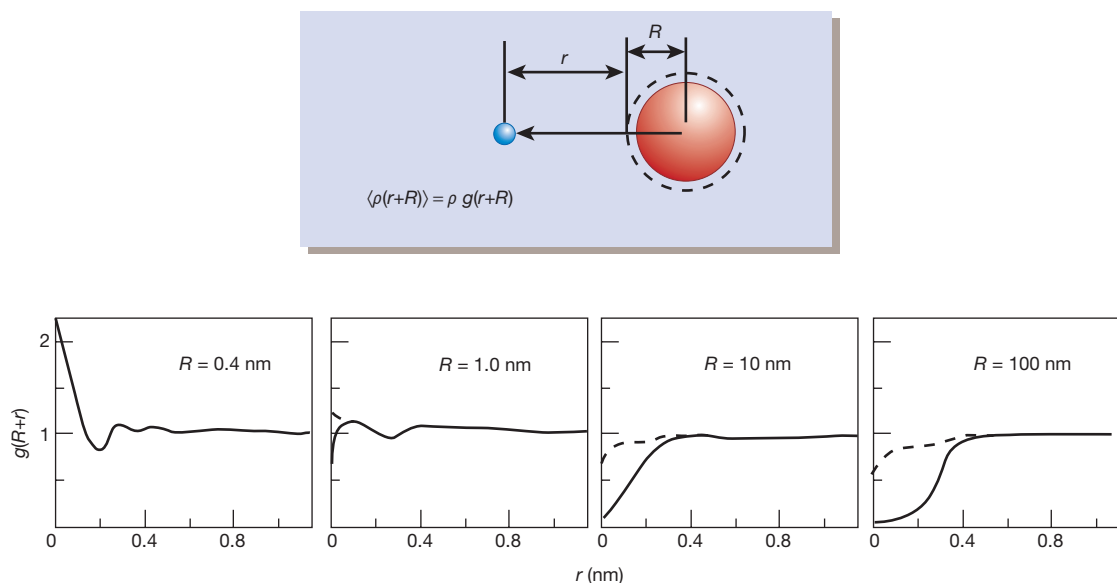


Figure 3 | The average equilibrium density of water a distance $r + R$ from spherical cavities in liquid water at standard conditions. R is the distance of closest approach between the centre of a water molecule (red circle) and the centre of the cavity (blue circle). The lines representing $g(r + R)$, the density, $\langle \rho(r + R) \rangle$ relative to that of the bulk water, ρ , are the results of microscopic theory^{21,24}. Solid lines refer to the ideal hydrophobic

solute, which expels water from the sphere of radius R . Dashed lines refer to the case where the cavity also interacts with water by means of a van der Waals attraction typical of that between water and a spherical cluster of oil. For cavities with radii less than 1 nm, the effects of this attraction on $g(R + r)$ are nearly negligible and not visible on the scale of the graph in the bottom panel.

and the liquid–vapour interface is especially clear for the largest of these cavities.

In formal statistical mechanical parlance, a wet surface is a surface covered by a macroscopically thick liquid layer or film, whereas a dry surface is covered by a macroscopically thick vapour layer. Capillary condensation is an example of wetting in this formal sense²⁷. In contrast, my use of these terms in this review refers to microscopic phenomena, and is descriptive rather than formal. In fact, van der Waals attraction between surface and solvent ensures that no natural macroscopic hydrophobic surface can be dry in the formal sense: the free-energy cost needed to move the liquid macroscopically far from the surface would be prohibitive. Nevertheless, compared with the density of water surrounding a small hydrophobic molecule or a hydrophilic surface, nanometre-sized and larger hydrophobic surfaces in water are indeed dewetted in the descriptive sense.

Experimental measures of the solvation free energy ΔG come from the free-energy change that occurs on transferring a hydrophobic molecule from its pure liquid to liquid water. For *n*-alkanes with 20 or fewer carbon atoms, the transfer free energy is a linear function of carbon number²⁸. Because the volume of a chain-like molecule grows linearly with the number of units, the linear trend in transfer free energies with carbon number is consistent with ΔG growing linearly with cavity volume, as expected for a small wet solute. However, the length of a 20-carbon alkane chain in coil conformation typically exceeds 1 nm, the crossover length beyond which large-solute solvation is expected. Because the curvatures of these hydrophobic surfaces are sufficiently high, they can still be hydrated as in the small-molecule regime. Globular conformations would present an extended hydrophobic surface with lower curvature that would prevent this type of hydration: if prevalent, these globular conformations would therefore lead to a change in the transfer free-energy trend. The fact that this trend remains linear for *n*-alkanes with 20, or fewer, carbons indicates that these molecules are rarely globular. In other words, with this number of carbons, hydrophobic forces seem to be insufficient to overwhelm chain entropy and stiffness that favour the coil state.

Driving force of assembly

The tendency for hydrophobic particles to cluster in water is readily understood in terms of the dependence of hydrophobic solvation on solute size. For example, imagine n identical small hydrophobic particles solvated in water, all well separated and thus solvated independently of each other. In this case, the overall solvation free energy is n times the solvation free energy for any one of the solutes, and it grows linearly with the overall excluded volume of the solutes.

When these n solutes cluster together to form a hydrophobic unit with an extended surface (that is, a surface with low curvature and larger than 1 nm²), the overall solvation free energy changes from growing linearly with solvated volume to growing linearly with solvated surface area. Figure 4 illustrates that if n is large enough, the solutes can form a cluster with a sufficiently large volume to surface ratio that its solvation free energy is lower than the overall solvation free energy of the individual solutes. This effect results in a favourable driving force for cluster assembly. The figure also illustrates that near ambient conditions, the driving force will get stronger with increasing T . This well-known trend is often cited as implying that hydrophobic interactions are entropic²⁹. Entropy does indeed contribute, but the assembly process is driven by the difference between the entropically dominated solvation free energy of small molecules and the enthalpically dominated solvation free energy of large surfaces.

A hydrophobic force that drives cluster assembly will be proportional to the change in solvated hydrophobic surface area only if all surfaces are sufficiently large. Generally, this is not the case. For example, when a collection of small hydrophobic units assembles into an extended cluster, such as that depicted in Fig. 4, the force driving the process will consist of one part that is proportional to the exposed surface area of the cluster, and another part that is proportional to the molecular volumes of the separate units. In physical situations such as this, with small as well as large length-scale regimes having a role, hydrophobic forces cease to be additive. Although hydrophobic forces between isolated small hydrophobic units can be decomposed into pair interactions, this is no longer possible as the units combine to form an extended hydrophobic surface and an associated solvent interface. This breakdown in additivity of hydrophobic forces is a manifest-

tation of the collective nature of interface formation.

At extremely low hydrophobic solute concentrations, the entropy of mixing solute molecules throughout a macroscopic container (which gives an effectively infinite volume) overwhelms the finite driving force to assemble, making clusters of finite size no more than metastable. On the other hand, finite solvent volumes will give rise to modest solute densities, and thus modest entropies of mixing, which need not defeat finite driving forces for hydrophobic association of the solutes. A computer simulation using such a finite solvent volume confirms that stable association requires hydrophobic clusters that extend beyond a critical radius of about 1 nm (ref. 30).

When hydrophobic particles are connected in a chain, the entropy favouring the extended state is finite, even in a system with infinite volume. Hydrophobic collapse in such a situation, with the chain going from a coil to a globule, involves transition states that coincide with the formation of critical nuclei of hydrophobic particles³¹; subsequent motions are committed to the stable globular conformations. The critical nuclei are formed in many ways, but always lead to the formation of a liquid–vapour-like interface. This interface then allows for fluctuations in water density, and these fluctuations permit the solvated chain to progress towards its stable globular state where the chain's interior is dry. Intriguingly, these solvent fluctuations, rather than chain dynamics, dominate the motion of the system through the transition states. The correct treatment of the hydrophobic collapse kinetics must therefore explicitly account for water dynamics, which precludes using solvent averaged potentials of mean force.

Role of weak attractive forces

Weak attractive forces, such as van der Waals interactions, are generally assumed to have little or no influence on the structure of a dense fluid³²: because density is high, the liquid is almost incompressible, eliminating all but small-length-scale density fluctuations. Instead, liquid structure is thought to be influenced only by packing effects, and by strong attractions that vary quickly in space, such as hydrogen bonds. This idea is a useful starting point for thinking about dense fluids. But it is not entirely applicable in the context of hydrophobicity, which involves interfaces forming in water near an extended hydrophobic surface.

Extended fluid interfaces near or at phase coexistence are often referred to as 'soft' because they can be translated in space with little or

Box 3 | Solvation effects of van der Waals interactions

In the case of small solutes, van der Waals attractions have little influence on the surrounding water structure (see also Fig. 3). Their contribution to the overall solvation free energy ΔG can therefore be estimated according to¹¹

$$\Delta G = \Delta G_v + \rho \int d\mathbf{r} g_v(\mathbf{r}) u(\mathbf{r}) \quad (6)$$

where ρ is the solute density, $g_v(\mathbf{r})$ the relative average density of solvent in the presence and absence of the solute cavity, and $u(\mathbf{r})$ the added van der Waals potential acting on the solvent at position $\mathbf{r}$. The integral in equation (6) is approximately linear in volume for small solutes, as is ΔG_v . The accuracy of equation (6) in estimating van der Waals contributions to solvation free energy has been verified by comparison against computer simulations⁶⁵.

The van der Waals contribution affects the overall solvation free energy ΔG appreciably, but its influence on the entropy of solvation ΔS ($= -\partial G/\partial T$) is rather subtle, mediated only through the temperature dependence of the cavity–water distribution function $g_v(\mathbf{r})$. This justifies using a cavity model to estimate hydrophobic solvation entropies and heat capacities^{50,52}.

Van der Waals interactions are attractive, so $u(\mathbf{r})$ is negative and contributes to ΔG such that it favours solvation. This contribution is primarily enthalpic, whereas ΔG_v itself is primarily entropic.

In the case of large hydrophobic solutes interacting through van der Waals forces with surrounding water, the average water density is well approximated as the bulk density at all points not occupied by the solute (see also Fig. 3). That is, $g_v(\mathbf{r})$ in equation (6) can be approximated by unity for all $\mathbf{r}$ outside the solute, and by zero otherwise. The integral then gives a favourable enthalpic contribution to ΔG , but scales linearly with solute surface area rather than with solute volume. For large solutes, therefore, size scaling and trends with temperature are the same for both the cavity and the van der Waals contributions to ΔG .

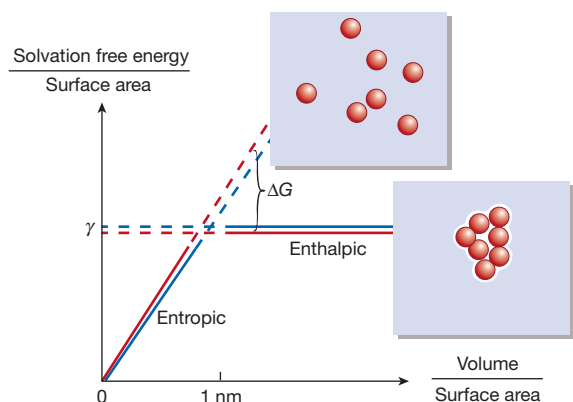


Figure 4 | The driving force, ΔG , for assembling a cluster of small hydrophobic particles. For large enough clusters, ΔG is a favourable driving force. The horizontal and sloping lines indicate the behaviour of the solvation free energy for the assembled and disassembled cluster, respectively. Red lines indicate the free energies at a higher liquid temperature; blue lines indicate the free energies at a lower temperature. The liquid–vapour surface tension is indicated by γ . 'Volume' and 'surface area' denote the volume excluded to water, and the solvated surface area of that volume, respectively.

no free-energy change¹. As a result of this softness, the location of the liquid–vapour-like interface, and the average liquid density near a large hydrophobic surface, can be significantly affected by van der Waals attractions. As illustrated in Fig. 3, for the case of a small hydrophobic solute where a soft interface does not accompany solvation, adding typical oil–water van der Waals attractions has essentially no effect on the average density of the surrounding water. But in the case of larger solutes that are accompanied by a soft interface, adding weak attractive forces brings the liquid interface into contact with the hydrophobic surface.

The resultant close proximity of the liquid interface to the hydrophobic surface and the larger average value of the liquid density near that surface has caused some to overlook dewetting, imagining the surface to be wet instead²⁴. A wet surface, however, is markedly different: although interfacial displacements and thus density fluctuations are substantial for water at its liquid–vapour interface and for water near an extended hydrophobic surface, they are largely absent for water adjacent to a wet surface. As a result of weak, but ubiquitous, attractive forces between oil and water, the main signature of dewetting is thus not the mean water density near the hydrophobic surface but the size of spontaneous density fluctuations. These fluctuations determine the degree to which weak attractions can affect the mean water density, even though the specific value of the mean for a particular choice of attraction has little physical significance. In contrast, the likelihood of fluctuations has an important role in the kinetics of hydrophobic assembly³¹.

Experiments confirm that the oil–water interface is much like the liquid–vapour interface of water, with no significant excess water density relative to the bulk water^{33–36}. In fact, the water–oil surface tension can be accurately approximated as the liquid–vapour surface tension plus the negative contribution arising from oil–water van der Waals interactions (see also Box 3). This readily explains why the tension of the oil–water interface is 20% lower than the surface tension of water²¹.

Similarly, the effect of van der Waals attractions on the solvation energies of small hydrophobic molecules can be readily computed, given

that these attractions have no appreciable effect on average solvent density (Box 3). This fact is codified in equation (6) in Box 3. Theory accounting for these attractions can also explain why methane, ethane and propane all have similar solvation free energies, an observation that is often incorrectly perceived as puzzling and incompatible with our understanding of hydrophobic interactions (see Box 4). Finally, because scaling with respect to solute size is not affected by van der Waals interactions in either the small-molecule or large-surface regime, it is straightforward to anticipate the effects of these interactions on the crossover length and on the thermodynamic driving force for clustering.

Strong attractive forces and amphiphile assembly

The effects of strong attractions between water and hydrophilic units differ from the effects of weak interactions, in that strong attractions tend to localize water molecules to specific locations and thereby limit fluctuations. This effect is important for the self-assembly in water of amphiphilic molecules, which contain hydrophobic as well as hydrophilic components. Such solutions form an array of mesoscopic assemblies that are at least partly stabilized by hydrophobic forces^{1,29,37}.

Box 4 | The small alkane solubility 'paradox'

In water at standard conditions, the solvation free energies of methane (CH_4), ethane ($\text{CH}_3\text{-CH}_3$), and propane ($\text{CH}_3\text{-CH}_2\text{-CH}_3$) are all about the same, differing by no more than 10% from 2 kcal mol⁻¹ (ref. 66). This similarity, often perceived as perplexing, is readily explained.

Consider first methane. Its solvation free energy ΔG_{Me} accounts for the formation of the methane cavity, $\Delta G_{\text{Me}}^{(0)}$, and for van der Waals attraction between methane and water, $-\epsilon$. We assume that methane forms a spherical cavity in water with $R \approx 0.35$ nm, and use equation (4) from Box 2 to estimate $\Delta G_{\text{Me}}^{(0)} \approx 7$ kcal mol⁻¹. From $\Delta G_{\text{Me}} = \Delta G_{\text{Me}}^{(0)} - \epsilon$, and the experimental value of ΔG_{Me} of about 2 kcal mol⁻¹, we obtain $\epsilon \approx 5$ kcal mol⁻¹.

The solvation free energy of ethane ΔG_{Et} might be estimated relative to that of two methane molecules, given that the solvation of an ethane molecule should be nearly identical to that of two overlapping methane particles⁶⁷. (The idea of two undistorted methane molecules occurring in such close proximity is unrealistic, but it is only the appearance of the solute to the solvent that determines the solvation free energy.) Thus $\Delta G_{\text{Et}} \approx 2\Delta G_{\text{Me}} + \Delta w(L)$, where $\Delta w(L)$ is the reversible work to bring two methane cavities in water to a separation L (equal to the C-C bond length in ethane). The reversible work must account for the difference in solvation free energies for cavities at infinite separation ∞ and at separation L , $\Delta G_{\text{v}(L)} - \Delta G_{\text{v}(\infty)}$, and for attractive methane-methane interactions, $u_{\text{MeMe}}(L)$. (The latter must be added to the cavity-cavity potential of mean force to account for the fact that the full radial distribution function is well approximated by the cavity radial distribution function³².) For water at standard conditions, the result is $\Delta w(L) = \Delta G_{\text{v}(L)} - \Delta G_{\text{v}(\infty)} + u_{\text{MeMe}}(L) \approx -2$ kcal mol⁻¹. Together with the methane solvation free energy ΔG_{Me} , this yields an ethane solvation free energy $\Delta G_{\text{Et}} \approx 2$ kcal mol⁻¹.

Similarly, the solvation free energy of propane ΔG_{Pr} might be estimated by considering separated cavities for methane and ethane in water that are reversibly moved together, so as to form a propane cavity. As this can be done roughly collinearly, the reversible work required to achieve this should be about the same as that needed for joining two methane cavities: $\Delta G_{\text{Pr}} \approx \Delta G_{\text{Et}} + \Delta G_{\text{Me}} + \Delta w(L)$. Again, $\Delta w(L)$ approximately cancels ΔG_{Me} , so ΔG_{Pr} has a value similar to ΔG_{Et} and ΔG_{Me} . Similar arguments can be used to estimate solvation free energies of short cyclo-alkanes in water relative to those of the corresponding normal alkane, indicating that $\Delta G_{\text{cyclo}} \approx \Delta G_{\text{normal}} + \Delta w(L)$. Experiment⁶⁸ confirms that the former is indeed between 1–2 kcal mol⁻¹ less than the latter.

The fact that $\Delta w(L)$ is close to $-\Delta G_{\text{Me}}$ is a coincidence, true for water at standard conditions but not true in general. Somewhat different values for the van der Waals attraction ϵ and the mean water density fluctuation χ_v would not change the basic physics, but would void the near cancellation that results in similar solvation free energies for methane, ethane and propane. The argument sketched here does, however, result in the generally valid prediction that the solvation free energy of a small normal alkane chain should scale linearly with the number of carbons, with a slope $\Delta G_{\text{Me}} + w(L)$. The scaling is found experimentally for the free energy to transfer an alkane chain from oil to water^{28,29}, with the cavity model yielding an estimate^{8,9} of the difference between $\Delta G_{\text{Me}} + w(L)$ for oil and for water that agrees well with the experimental slope.

The principles that hold for purely hydrophobic solutes also apply to molecules containing some hydrophilic units. But additional entropic effects arise because molecular configurations are restricted when both hydrophilic and hydrophobic interactions need to be accommodated. This is illustrated in the simplest example of amphiphilic assembly — micelle formation (see Fig. 5).

The free energy for solvating an amphiphile either in water or in a micelle with an oily interior can be approximated by considering the free-energy contributions of its hydrophilic head and hydrophobic tail separately. The hydrophilic head will always be in an aqueous environment, so the free energy for transferring an amphiphile from a micelle into water, g_{trans} , approximately equals the free energy for transferring the corresponding hydrophobic molecule from oil to water. The force driving the assembly of a micelle with n amphiphilic molecules, ΔG_n , will therefore contain the contribution $-ng_{\text{trans}}$. This contribution favours the formation of a micelle, but is opposed by two other free-energy contributions. One arises from the fact that the formation of a micelle involves the creation of an interface, as is familiar from nucleation theory¹². A second opposing contribution is entropic. This accounts for the reduction in configurations available to amphiphilic molecules when hydrophobic chains are constrained to lie within a micelle while the head groups are confined to the micelle surface. This effect limits micelle growth to a finite size: for large n , there is simply no space available to maintain a dense micelle interior while simultaneously placing head groups on the exterior.

Given these different free-energy contributions, the principle of association equilibrium¹¹ (that is, the law of mass action) readily explains³⁸ why the critical micelle concentration decreases exponentially with amphiphile chain length, and further why this concentration decreases with increasing T at low T , but increases with T at high T . The exponential behaviour with chain length reflects the linearity of oil–water transfer free energy as a function of hydrophobic molecule chain length²⁹. The non-monotonic trend with temperature reflects the fact that transfer free energies are similarly non-monotonic (that is, ΔS for hydrophobic solvation passes through zero at a convergence temperature).

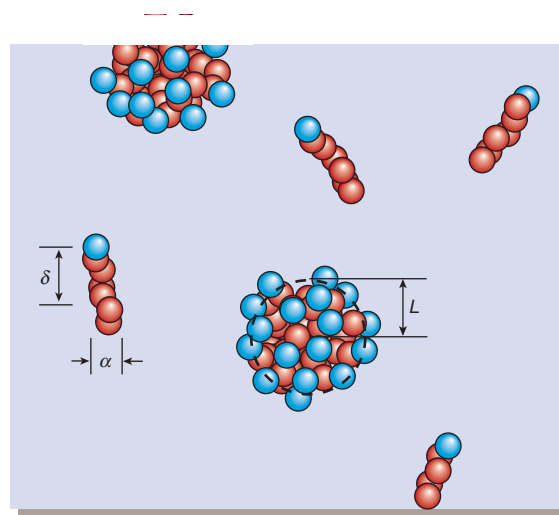


Figure 5 | Length scales of amphiphiles in dynamic equilibrium with micelles. The blue and red spheres depict the hydrophilic heads and the hydrophobic tails, respectively, of the amphiphiles. The typical length over which hydrophobic and hydrophilic components are separated within a single molecule is given by δ . Assuming a roughly spherical structure and tightly packed oily components in the centre, the micelle radius is $L \approx (\alpha^2 \delta^2 n)^{1/3}$, where n is the number of surfactants in the micelle. Accounting for the hydrophobic driving force for assembly, plus interfacial and entropic free-energy costs, the average size of the micelle is found to be $n \approx \beta \gamma \delta^2$ (where $\beta = 1/k_B T$ and γ is the oil–water surface tension), and the critical micelle concentration approximately $(1/\alpha^3) \exp[-\beta g_{\text{trans}} + c(\beta \gamma \delta^2)^{2/3}]$, where g_{trans} is the oil-to-water transfer free energy for the oily chain of an amphiphile and $c = (5,832/49)^{1/3} \approx 4.9$ (ref. 38).

Is water special?

Much has been written on the question of whether water is special and about its importance in biology (for examples, see refs 39, 40). Considering its thermodynamic anomalies, patterns of hydrogen bonding, ability to support fast proton transport and so forth, liquid water is obviously a unique solvent. Nevertheless, the two physical features responsible for hydrophobicity — the solvent being close to phase coexistence with its vapour, and solute–solvent interactions being significantly less attractive than solvent–solvent interactions — are not particularly unusual, at least in an abstract sense.

The proximity of phase coexistence and the imbalance of attractive forces can be satisfied with model systems that have little in common with water. Their only similarity with water is a high molecular density at conditions of low pressure; that is, the solvent, like water, is close to its triple point. For example, a Lennard–Jones solvent^{41–43}, a two-dimensional fluid model⁴⁴ and even a two-state lattice gas^{45,46} all provide perfectly acceptable models that will show behaviour akin to hydrophobicity, including amphiphilic self assembly^{47–49}.

In nature, however, it is difficult to find a pairing of liquids that show an attractive force imbalance that is similar to that found for water and oil. This imbalance is due to hydrogen bonding between water molecules, and produces the crossover in solvation behaviour at a length scale of 1 nm and at thermodynamic conditions near 1 atm pressure and room temperature. Furthermore, the temperature dependence of the structure of water's hydrogen-bond network leads to the relatively large temperature dependence of solvation entropies described earlier. This temperature dependence of solvation entropies has a significant role in the temperature dependence of assembly processes (see Fig. 4), and is considered important in the context of protein folding^{50–53}.

To the future

It is one thing to understand the forces underlying hydrophobic interactions, but quite another to appreciate all the interesting implications such forces have. For example, a correct and useful theory of hydrophobic effects should provide quantitative guidance for the study of biophysical systems. Do we now have such a theory? Perhaps, but its usefulness remains to be demonstrated.

One area where progress might soon be made is the study of protein–protein interactions controlling self-assembly of large protein complexes. Here, challenges arise from the presence of two types of disorder⁵⁴: the fact that the hydrophobic surfaces of proteins are laced with hydrophilic units, and that the surfaces are irregular in shape. These features will affect the size of hydrophobic units, and hence how hydrophobic interactions arise, as well as their strengths and kinetics. For example, hydrophilic units will probably lessen the extent of drying and thereby allow water to lubricate the final stages of assembly in a protein complex⁵⁵. In addition, the exact placement of hydrophilic units and the specifics of surface topography should influence the complementarities of protein surfaces. Similar issues should also present important challenges when trying to elucidate the role of hydrophobic effects in intramolecular processes such as protein folding.

Materials science is another area where our understanding of hydrophobic effects might prove useful. For example, drying-mediated self-assembly of nanoparticles involves competition between the kinetics of evaporation and the time scales with which solvated nanoparticles diffuse on a substrate⁵⁶. How these effects might be enriched by surfaces that nucleate evaporation remains to be explored. Such systems, and also the adhesion between macroscopic surfaces, involve pertinent length scales other than just the crossover length (which is essentially the radius of a critical nucleus of oil in a super-saturated water–oil mixture). For example, the confinement of water by two macroscopic hydrophobic surfaces in solution can induce evaporation of the confined liquid because the oil–water interfacial free energy outweighs the free energy favouring liquid water over its vapour phase. For flat, parallel surfaces in water at standard conditions, the inter-surface separation exceeds 1,000 nm where evaporation becomes favourable.

Surface-induced evaporation of this sort between macroscopic

plates would produce a macroscopic adhesive force, but it requires that evaporation be kinetically accessible. This can occur through interfacial fluctuations that bring two separate interfaces in water into contact⁵⁷. In the case of hydrated mesoscopic hydrophobic surfaces, such fluctuations are feasible because the confined liquid remains stable until the surfaces are relatively close together. But for macroscopic surfaces, the free-energy cost for such fluctuations is prohibitive, and evaporation is possible only through other pathways. In particular, a liquid sufficiently confined by unfavourable surfaces will become mechanically unstable, causing, in effect, spinodal decomposition. For pure water, this instability is estimated²⁰ to occur when two macroscopic hydrophobic surfaces are separated by about 5 nm, which is similar to the separation at which forces become unstable for two hydrated hydrophobic surfaces in a surface-force apparatus⁵⁸.

Whether these specific phenomena prove significant, there is no doubt that the hydrophobic effect is fundamental. The varied and possible complexities of aqueous solutions where the hydrophobic effect is manifested provide ample opportunity to probe the underlying general principles outlined in this review. ■

1. Safran, S. A. *Statistical Thermodynamics of Surfaces, Interfaces and Membranes* Ch. 1–3 (Addison-Wesley, Reading, 1994).
2. Tanford, C. The hydrophobic effect and the organization of living matter. *Science* **200**, 1012–1018 (1978).
3. Kauzmann, W. Some forces in the interpretation of protein denaturation. *Adv. Prot. Chem.* **14**, 1–63 (1959).
4. Tanford, C. & Reynolds, J. *Nature's Robots: A History of Proteins* Ch. 12 (Oxford Univ. Press, Oxford, 2001).
5. Tanford, C. How protein chemists learned about the hydrophobic factor. *Protein Sci.* **6**, 1358–1366 (1997).
6. Dixh, S., Crain, J., Pooh, W. C. K., Finney, J. L. & Soper, A. K. Molecular segregation observed in a concentrated alcohol–water solution. *Nature* **416**, 829–832 (2002).
7. Swope, W. C. & Andersen, H. C. A molecular dynamics method for calculating the solubility of gases in liquids and the hydrophobic hydration of inert gas atoms in aqueous solution. *J. Phys. Chem.* **88**, 6548–6556 (1984).
8. Pratt, L. R. & Chandler, D. Theory of the hydrophobic effect. *J. Chem. Phys.* **67**, 3683–3704 (1977).
9. Pratt, L. R. & Chandler, D. Hydrophobic solvation of nonspherical solutes. *J. Chem. Phys.* **73**, 3430–3433 (1980).
10. Pratt, L. R. & Chandler, D. Hydrophobic interactions and osmotic second virial coefficients for methanol in water. *J. Solution Chem.* **9**, 1–17 (1980).
11. Chandler, D. *Introduction to Modern Statistical Mechanics* Ch. 3, 5–7 (Oxford Univ. Press, New York, 1987).
12. Barrat, J.-L. & Hansen, J.-P. *Basic Concepts for Simple and Complex Liquids* (Cambridge Univ. Press, Cambridge, 2003).
13. Sharp, K. A., Nicholls, A., Fine, R. & Honig, B. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science* **252**, 106–109 (1991).
14. Ashbaugh, H. S., Kaler, E. W. & Paulaitis, M. E. A “Universal” surface area correlation for molecular hydrophobic phenomena. *J. Am. Chem. Soc.* **121**, 9243–9244 (1999).
15. Tanford, C. Interfacial free energy and the hydrophobic effect. *Proc. Natl Acad. Sci. USA* **76**, 4175–4176 (1979).
16. Bowron, D. T., Filipponi, A., Roberts, M. A. & Finney, J. L. Hydrophobic hydration and the formation of a clathrate hydrate. *Phys. Rev. Lett.* **81**, 4164–4167 (1998).
17. Stillinger, F. H. Structure in aqueous solutions of nonpolar solutes from the standpoint of scaled-particle theory. *J. Solution Chem.* **2**, 141–158 (1973).
18. Dill, K. A. & Bromberg, S. *Molecular Driving Forces* Ch. 16, 30 (Garland Science, New York, 2003).
19. Murphy, K. P. Hydration and convergence temperatures: on the use and interpretation of correlation plots. *Biophys. Chem.* **51**, 311–326 (1994).
20. Lum, K., Chandler, D. & Weeks, J. D. Hydrophobicity at small and large length scales. *J. Phys. Chem. B* **103**, 4570–4577 (1999).
21. Huang, D. M. & Chandler, D. The hydrophobic effect and the influence of solute–solvent attractions. *J. Phys. Chem. B* **106**, 2047–2053 (2002).
22. Wallqvist, A. & Berne, B. J. Computer simulation of hydrophobic hydration forces on stacked plates at short range. *J. Phys. Chem.* **99**, 2893–2899 (1995).
23. Gerstein, M. & Lynden-Bell, R. M. Simulation of water around a model protein helix. 1. Two-dimensional projections of solvent structure. *J. Phys. Chem.* **97**, 2982–2990 (1993).
24. Ashbaugh, H. S. & Paulaitis, M. E. Effect of solute size and solute–water attractive interactions on hydration water structure around hydrophobic solutes. *J. Am. Chem. Soc.* **123**, 10721–10728 (2001).
25. Huang, D. M., Geissler, P. L. & Chandler, D. Scaling of hydrophobic free energies. *J. Phys. Chem. B* **105**, 6704–6709 (2001).
26. Lee, C. Y., McCammon, J. A. & Rossky, P. J. The structure of liquid water at an extended hydrophobic surface. *J. Chem. Phys.* **80**, 4448–4455 (1984).
27. Rowlinson, J. S. & Widom, B. *Molecular Theory of Capillarity* Ch. 8 (Oxford Univ. Press, Oxford, 1982).
28. Smith, R. & Tanford, C. Hydrophobicity of long chain n-alkyl carboxylic acids, as measured by their distribution between heptane and aqueous solutions. *Proc. Natl Acad. Sci. USA* **70**, 289–293 (1973).
29. Tanford, C. *The Hydrophobic Effect: Formation of Micelles & Biological Membranes* Ch. 1–4, 8 (Wiley, New York, 1980).
30. Raschke, T. M., Tsai, J. & Levitt, M. Quantification of the hydrophobic interaction by simulations of the aggregation of small hydrophobic solutes in water. *Proc. Natl Acad. Sci. USA* **98**, 5965–5969 (2001).

31. TenWolde, P. R. & Chandler, D. Drying induced hydrophobic polymer collapse. *Proc. Natl Acad. Sci. USA* **99**, 6539–6543 (2002).
32. Chandler, D., Weeks, J. D. & Andersen, H. C. The van der Waals picture of liquids, solids and phase transformations. *Science* **220**, 787–794 (1983).
33. Scatena, L. F. & Richmond, G. L. Orientation, hydrogen bonding, and penetration of water at the organic/water interface. *J. Phys. Chem. B* **105**, 11240–11250 (2001).
34. Ball, P. Chemical physics: how to keep dry in water. *Nature* **423**, 25–26 (2003).
35. Jensen, T. R. et al. Water in contact with extended hydrophobic surfaces: Direct evidence of weak dewetting. *Phys. Rev. Lett.* **90**, 086101.1–086101.4 (2003).
36. Zhang, X., Zhu, Y. & Granick, S. Hydrophobicity at a janus interface. *Science* **295**, 663–666 (2002).
37. Larson, R. G. *Structure and Rheology of Complex Fluids* Ch. 12 (Oxford Univ. Press, New York, 1999).
38. Maibaum, L., Dinner, A. R. & Chandler, D. Micelle formation and the hydrophobic effect. *J. Phys. Chem. B* **108**, 6778–6781 (2004).
39. Finney, J. L. Water? What's so special about it? *Phil. Trans. R. Soc. Lond. B* **359**, 1145–1165 (2004).
40. Tanford, C. *Ben Franklin Stilled the Waves* Ch. 14 (Oxford Univ. Press, Oxford, 2004).
41. Huang, D. M. & Chandler, D. Cavity formation and the drying transition in a Lennard-Jones fluid. *Phys. Rev. E* **61**, 1501–1506 (2000).
42. Bolhuis, P. G. & Chandler, D. Transition path sampling of cavitation between molecular scale solvophobic surfaces. *J. Chem. Phys.* **113**, 8154–8160 (2000).
43. Weeks, J. D. Connecting local structure to interface formation: A molecular scale van der Waals theory of non-uniform liquids. *Annu. Rev. Phys. Chem.* **53**, 533–562 (2002).
44. Southall, N. T. & Dill, K. A. The mechanism of hydrophobic solvation depends on solute radius. *J. Phys. Chem. B* **104**, 1326–1331 (2003).
45. Binder, K., Landau, D. P. & Ferrenberg, A. M. Wetting and capillary condensation of lattice gases in thin film geometry. *Ber. Bunsenges. Phys. Chem.* **98**, 340–345 (1994).
46. Maibaum, L. & Chandler, D. A coarse-grained model of water confined in a hydrophobic tube. *J. Phys. Chem. B* **107**, 1189–1193 (2003).
47. Smit, B. et al. Computer simulations of a water/oil interface in the presence of micelles. *Nature* **348**, 624–625 (1990).
48. Wu, D., Chandler, D. & Smit, B. Electrostatic analogy for surfactant assemblies. *J. Phys. Chem.* **96**, 4077–4083 (1992).
49. Larson, R. G. Monte Carlo simulation of microstructural transitions in surfactant systems. *J. Chem. Phys.* **96**, 7904–7918 (1992).
50. Garde, S., Hummer, G., Garcia, A. E., Paulaitis, M. E. & Pratt, L. R. Origin of entropy convergence in hydrophobic hydration and protein folding. *Phys. Rev. Lett.* **77**, 4966–4968 (1998).
51. Murphy, K. P., Privalov, P. L. & Gill, S. J. Common features of protein unfolding and dissolution of hydrophobic compounds. *Science* **247**, 559–561 (1990).
52. Huang, D. & Chandler, D. Temperature and length scale dependence of hydrophobic effects and their possible implications for protein folding. *Proc. Natl Acad. Sci. USA* **97**, 8324–8327 (2000).
53. Baldwin, R. L. Temperature dependence of the hydrophobic interaction in protein folding. *Proc. Natl Acad. Sci. USA* **83**, 8069–8072 (1986).
54. Gerstein, M. & Chothia, C. Packing at the protein-water interface. *Proc. Natl Acad. Sci. USA* **93**, 10167–10172 (1996).
55. Levy, Y. & Onuchic, J. N. Water and proteins: A love-hate relationship. *Proc. Natl Acad. Sci. USA* **101**, 3325–3326 (2004).
56. Rabani, E., Reichman, D. R., Geissler, P. L. & Brus, L. E. Drying-mediated self-assembly of nanoparticles. *Nature* **426**, 271–274 (2003).
57. Lum, K. & Chandler, D. Phase diagram and free energies of vapor films and tubes for a confined fluid. *Int. J. Thermophys.* **19**, 845–855 (1998).
58. Parker, J. L., Claesson, P. M. & Attard, P. Bubbles, cavities, and the long-ranged attraction between hydrophobic surfaces. *J. Phys. Chem.* **98**, 8468–8480 (1994).
59. Frenkel, D. & Smit, B. *Understanding Molecular Simulation* 2nd edn Ch. 7 (Academic, San Diego, 2002).
60. Hummer, G., Garde, S., Garcia, A. E., Pohorille, A. & Pratt, L. R. An information theory model of hydrophobic interactions. *Proc. Natl Acad. Sci. USA* **93**, 8951–8955 (1996).
61. Morrison, T. J. & Billett, F. The salting-out of non-electrolytes. Part II. The effect of variation in non-electrolytes. *J. Chem. Soc.* 3819–3822 (1952).
62. Chandler, D. Gaussian field model of fluids with an application to polymeric fluids. *Phys. Rev. E* **48**, 2898–2905 (1993).
63. Silverstein, K. A. T., Haymet, A. D. J. & Dill, K. A. The strength of hydrogen bonds in liquid water and around nonpolar solutes. *J. Am. Chem. Soc.* **122**, 8037–8041 (2000).
64. Gallagher, K. R. & Sharp, K. A. A new angle on heat capacity changes in hydrophobic solvation. *J. Am. Chem. Soc.* **125**, 9853–9860 (2003).
65. Gallicchio, E., Kubo, M. M. & Levy, R. M. Enthalpy-entropy and cavity decomposition of alkane hydration free energies: numerical results and implications for theories of hydrophobic solvation. *J. Phys. Chem. B* **104**, 6271–6285 (2000).
66. Mackay, D. & Shiu, W. Y. A critical review of Henry's Law constants for chemicals of environmental interest. *J. Phys. Chem. Ref. Data* **10**, 1175–1199 (1981).
67. Ben-Naim, B. *Hydrophobic Interactions* Ch. 3 (Plenum, New York, 1980).
68. Mackay, D., Shiu, W. Y. & Ma, K. C. *Illustrated Handbook of Physical-Chemical Properties and Environmental Fate of Organic Chemicals* Vol. III, 117, 170 (CRC Press, Boca Raton, 1993).

Acknowledgements My research on hydrophobicity was supported by the National Science Foundation and the US Department of Energy. I thank L. Maibaum for discussions and help in writing this review. I am also indebted to R. Levy and C. Tanford for constructive criticisms of an earlier draft. My choices for references are illustrative, not comprehensive.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to D.C. (chandler@cchem.berkeley.edu).

Controlled microfluidic interfaces

Javier Atencia¹ & David J. Beebe¹

The microfabrication technologies of the semiconductor industry have made it possible to integrate increasingly complex electronic and mechanical functions, providing us with ever smaller, cheaper and smarter sensors and devices. These technologies have also spawned microfluidics systems for containing and controlling fluid at the micrometre scale, where the increasing importance of viscosity and surface tension profoundly affects fluid behaviour. It is this confluence of available microscale engineering and scale-dependence of fluid behaviour that has revolutionized our ability to precisely control fluid/fluid interfaces for use in fields ranging from materials processing and analytical chemistry to biology and medicine.

The effects of gravity and inertia dominate our experiences of the physical world. But as systems are reduced in size, phenomena such as diffusion, surface tension and viscosity become ever more important; at the micrometre scale they can dominate and result in a world that operates very differently from the macroscopic world we perceive and live in¹. Purcell provided a fascinating peek into such a world populated by microorganisms², showing that *Escherichia coli*, with a size of about 2 μm , moves more slowly than diffusing nutrients and waste. This means that rather than actively search for its food, *E. coli* can forage just as efficiently by simply waiting for food to diffuse past. There are plenty of other processes where nature uses the micrometre scale to its advantage. For instance, gas exchange occurs with relatively slow rates within our lungs through diffusion, but nevertheless is efficient overall because it can take place over a large total surface area of about 80 m^2 provided by large numbers of very small air spaces (the alveoli). Similarly, muscle contraction is triggered by calcium ion diffusion, yet large muscles are often activated very rapidly. In this case, the muscle fibres are highly ordered and consist of micrometre-sized repeating structural units; this design keeps diffusion distances short so that the ions can rapidly reach their target destination.

Unlike nature, we are only just beginning to harness microscale phenomena for practical use. This contrasts with our understanding of the behaviour of particles and fluids at the microscale, which has a long history that can be traced back to capillary experiments by Hagen and Poiseuille in the middle of the nineteenth century^{3,4}. Navier and Stokes provided important contributions to fluid dynamic theory in the beginning of the nineteenth century^{5,6}, and Taylor extended the field with his studies of diffusion under laminar flow in the 1950s⁷. But practical, creative use of this knowledge has had to await the availability of technologies for building microscale systems in a controlled and repeatable manner. That started to happen in the early 1980s, which saw the emergence of micro-electromechanical systems⁸ (MEMS). MEMS aimed to integrate electronics and mechanical elements such as sensors and actuators on a common substrate, by adapting the advanced fabrication capabilities of the microelectronics industry. The same fabrication technologies were subsequently also used to create devices for containing and controlling fluid at the micrometre scale^{9,10}, giving rise to the field of microfluidics. Much of the original motivation for microfluidics arose out of developments in biology that call for the ability to manipulate fluids on the cellular length scale, and by the desire to provide cheap and efficient diagnostic tools that require only small sample volumes¹¹. Microfluidic systems have now been

improved to the state where they are commercially available for biomolecular separations (Caliper Life Sciences) and emerging as promising tools for high-throughput discovery and screening studies in chemistry and materials science^{12,13}. But beyond the manipulation of liquids as such, microfluidic systems also let us exploit the scale-dependence of interface properties to develop a wide range of other applications, as we aim to illustrate here.

A striking demonstration of the potential for exquisite control of liquid interfaces at the micrometre scale appeared in 1992: a drop of water moves autonomously uphill when placed on a smooth surface that is treated so as to have a gradient in hydrophobicity¹⁴. Since then, many successful approaches have been developed for adjusting substrate sur-

Box 1 | The importance of scale

Reynolds number

The Reynolds number relates the ratio of inertial to viscous forces. Viscosity, the internal friction of a fluid, produces a resistance to shear and a tendency for the fluid to move in parallel layers known as laminar flow; and inertia, the tendency of a body in motion to retain its initial motion, counters laminar flow and can ultimately result in turbulent flow. Quantitatively, the Reynolds number is calculated as $Re = av/\nu$, where v is the velocity scale of the fluid, a is a characteristic distance of the system (in the case of flow through a pipe, a would be the pipe diameter), and ν is the kinematic viscosity of the fluid.

Peclet number

The Peclet number, Pe , provides an indication of the relative importance of diffusion and convection, diffusion being the random thermal motion of molecules within their surrounding environment and convection the transport as a result of bulk motion of a fluid²³. The Peclet number is defined as $Pe = U_s H/D$, where U_s is the average velocity of the flow, H is a characteristic length of the system perpendicular to the direction of the flow and D is the diffusion coefficient of the particle or molecule of interest.

Capillary number

The ratio between viscous and capillary forces is given by the capillary number. Capillarity is the rise or depression of a liquid in a small passage, such as a thin tube. Water inside a glass capillary tube will have a concave meniscus that is in equilibrium because of a pressure difference across the interface. Such a pressure difference exists whenever a liquid surface is curved (as in the case of liquid drops or soap bubbles²⁷), with the higher pressure found on the inner side of the curve. The capillary number is given by $Ca = U\mu/\gamma$, where U is the velocity of the flow, μ is the viscosity of the fluid, and γ is the surface tension.

¹Department of Biomedical Engineering, University of Wisconsin-Madison, 1550 Engineering Drive, Rm 2142 ECB, Madison, Wisconsin WI 53706, USA.

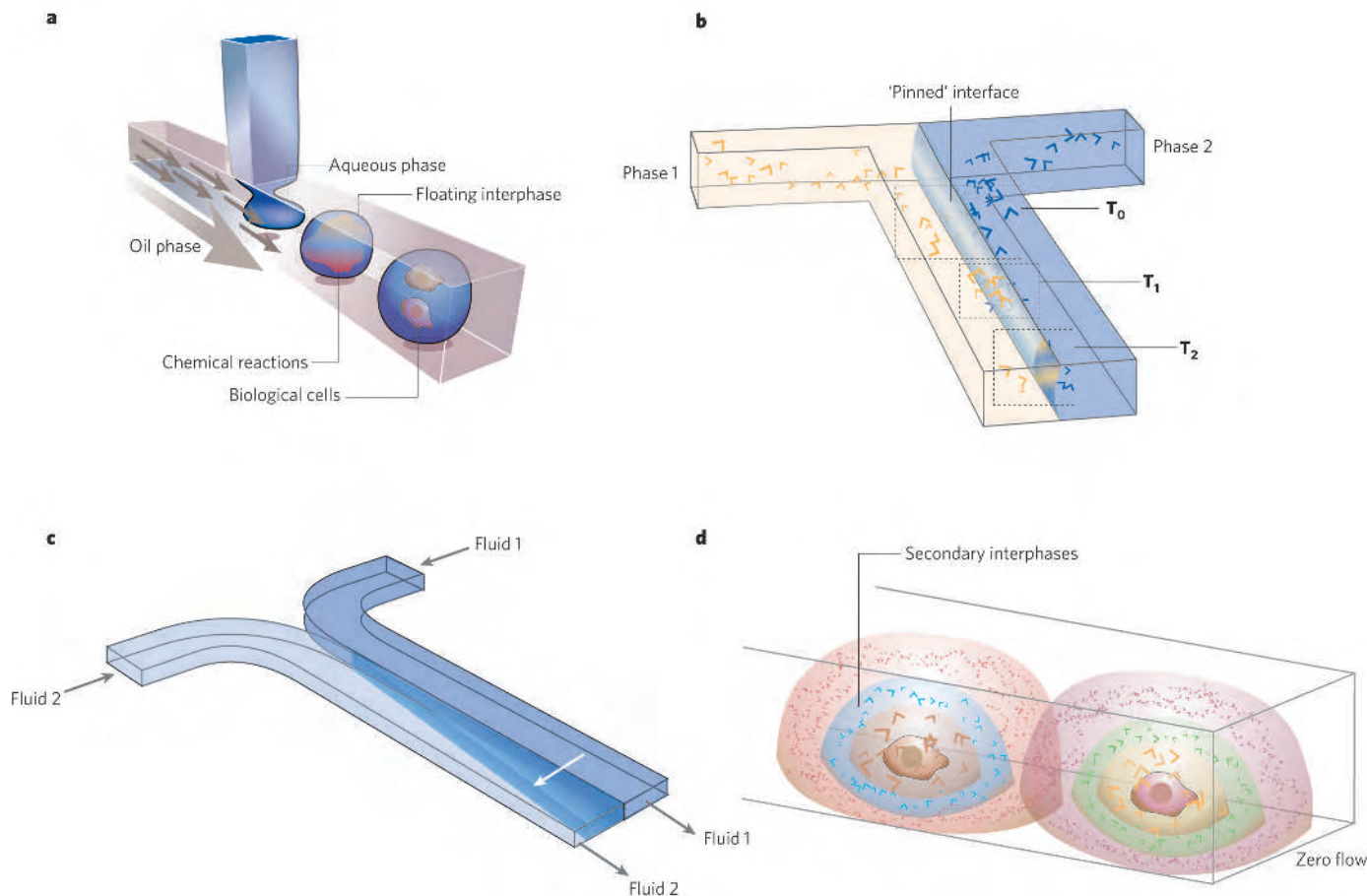


Figure 1 | Microfluidic interfaces provide unique functionality. **a**, Floating interfaces between immiscible fluids can be used to produce droplets of precise shape and varying content. These can act as microscale containers with permeable walls for performing and analysing reactions, creating custom magnetic or protein-coated vesicles, or transporting cargo. **b**, Pinned interfaces between immiscible liquids are created by selective surface patterning of a microchannel. They can be used to create vertical interfaces between liquid and air that, defying gravity, can hold liquids without breaking, or to stabilize the interface between immiscible fluids

allowing interfacial chemistry to form real walls. **c**, Moving interfaces between miscible liquids are created under laminar flow conditions between two streams of fluid flowing together because they do not mix except by diffusion, creating a diffusive interface with predictable geometry. **d**, Secondary interfaces arise in microscale channels because transport is mainly due to diffusion (convection-free environments). Diffusion can create complex but predictable patterns (interfaces) of solutes based on the diffusivity of the solutes.

face properties as a means of manipulating liquid drops, and sophisticated methods capable of controlling surface properties both temporally and spatially are now at our disposal^{15–18}. One such method — known as electrowetting — uses electrical control of contact angle to manipulate liquid droplets in real time¹⁹. This control capability can be used in digital microfluidics, the processing of discrete fluid packets that is of interest for the development of clinical diagnostic assays^{20,21}. But it is the ability of microfluidics to harness interfaces that is continuing to open new avenues of inquiry and application, and is the focus of this review.

In our discussions, we will go beyond the classical view of an 'interface' as the thin boundary layer that separates two distinct phases of matter (each of which may be a solid, a liquid or a gas) and that has properties distinct from those of the bulk material on either side. In addition to such classical interfaces, we also consider *de facto* interfaces such as the diffusive layers that appear if miscible fluids are brought into contact or a solute source is placed in a fluid²². A common theme is the precise control that microfluidics offers over the interface, permitting many applications that were not previously possible.

Fluid at the microscale

An obvious effect of shrinking a system to the micrometre scale is the huge increase in surface area relative to volume, often by several orders

of magnitude. For a fluid, the effect allows for more efficient mass and heat transfer in microsystems: relatively more interface is available for transfer to occur, and less total mass or energy needs to be transferred to reach the final state. Therefore both the creation and the homogenization of solute or temperature gradients are faster as system size is reduced.

Fluid behaviour in reduced dimensions will also be increasingly influenced by viscosity rather than inertia. In the case of microfluidic systems with simple geometries, this results in laminar flow. (Such behaviour occurs if the Reynolds number Re , which gives the ratio of viscous to inertial forces, is small; see Box 1). In laminar flow, diffusion can be effective for moving and mixing solutes on micrometre length scales. The relative importance of diffusion and convective bulk flow for transporting solute and solvent molecules is given by the Peclet number Pe (see Box 1), and can be readily adjusted through the choice of flow velocity and the dimensions of the system used.

The large ratio of surface area to volume typical for microfluidic systems ensures that surface tension can profoundly influence fluid behaviour. If surface tension varies along a surface or interface as a result of thermal or concentration gradients, for example, so-called Marangoni flows^{23,24} can arise and effectively homogenize the thermal or concentration gradients; the convective flows may even be used to

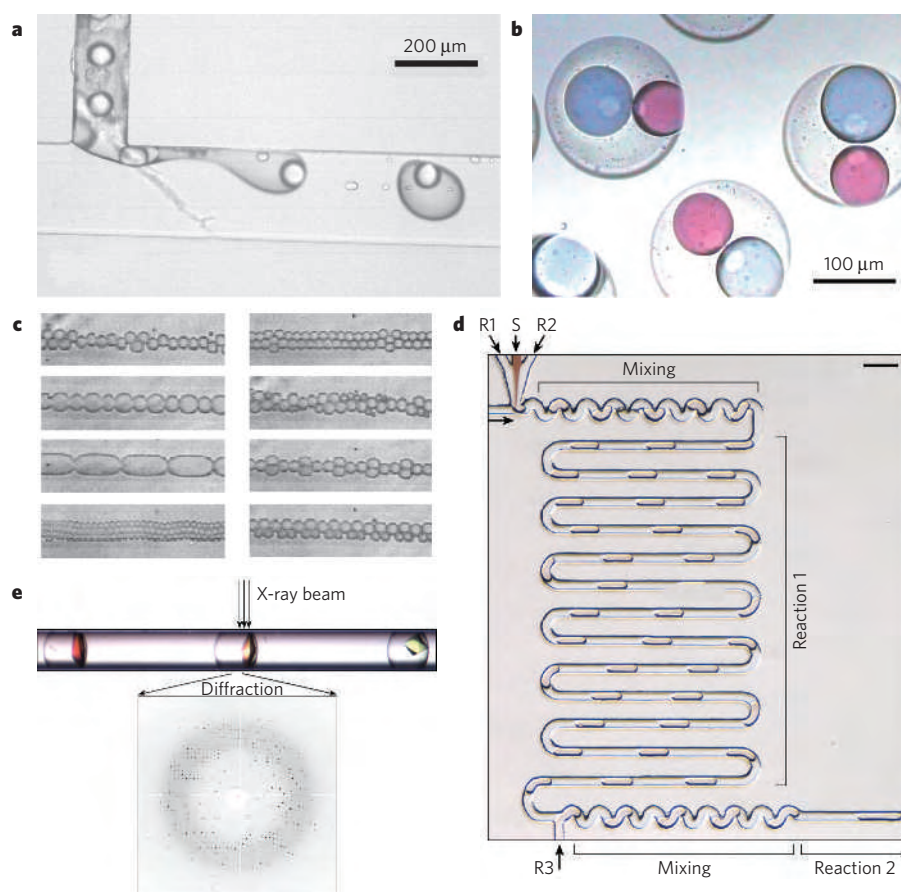


Figure 2 | Droplets as containers. Microfluidic technology allows droplets to be generated that can be used as containers in which different kind of reactions can be carried out in a controlled way. **a, b,** A T-junction is used for generating monodisperse double emulsions, with the single emulsion produced by another T-junction upstream with opposite wettability properties (hydrophobic/hydrophilic). Potential applications include encapsulation of therapeutic agents for targeted drug delivery and extractions across the thin layer separating the internal droplets. Reprinted with permission from ref. 34. Copyright (2003) American Chemical Society. **c,** The competition between capillary and viscous forces in a T-junction is used to generate three-dimensional patterns ('necklaces' and zigzag, for example). The goal is to study how systems operating far from equilibrium can produce regular patterns. Reprinted with permission from ref. 33.

Copyright (2001) American Physical Society. **d,** A microfluidic platform for performing a two-step reaction in which droplets are used as containers. Aqueous reagents R1 and R2 are merged in a T-junction to form a droplet immersed in oil. Mixing is aided by chaotic advection generated by shear stresses on the interface of the droplets, as they flow through a microchannel with alternating curves. After mixing the droplets flow through a longer channel to allow the reaction to proceed. At the end of the channel, another T-junction permits the injection of a third reagent R3, and the process can be repeated as desired. Reproduced with permission from ref. 44. Copyright (2004) Royal Society of Chemistry. **e,** Protein crystals are formed inside droplets in a glass capillary. The diffraction pattern of the crystal can be obtained directly without removing the droplets from the capillary. Reprinted with permission from ref. 45. Copyright (2003) Wiley-VCH.

move droplets and pump fluid^{25,26}. But the most significant surface tension phenomenon is probably capillarity — the rise or depression of a liquid in a small passage, driven by capillary forces which according to the Young–Laplace equation²⁷ become more significant relative to other forces such as gravity as a system's size is reduced²⁸. Surface properties can be selected to influence the competition between viscous forces and capillary forces. This makes it possible to adjust the competition (which is quantified by the capillary number Ca ; see Box 1) to control the generation, break-up and coalescence of droplets.

Comprehensive and quantitative reviews of fluid behaviour at the microscale are provided elsewhere^{29,30}, but the brief summary above already shows that it can differ markedly from that seen on the macroscale.

Our aim here is to illustrate that microfluidic systems offer a wealth of ways to exploit this unique behaviour — to create and control interfaces and make use of their interface properties. We will focus on interfaces between fluids (liquids and gas). In the case of immiscible fluids we consider droplets of one liquid dispersed in another, where the 'floating interface' between the two phases can act as a semipermeable container wall (Fig. 1a). Immiscible fluids flowing next to each other near a solid surface can also be separated by stable 'pinned interfaces',

which are maintained by the action of capillary forces and may act as membranes (Fig. 1b). If the fluids are miscible, there is clearly no defined interface: as the fluids are brought into contact they will mix over time, ultimately yielding a homogeneous fluid. But under laminar flow conditions, the boundary between two miscible fluids moving next to each other and mixing only through diffusion can be regarded as a dynamic or 'moving interface' that can be manipulated and put to practical use (Fig. 1c). Similarly, the diffusive layer forming around a stationary solute source, referred to here as a 'secondary interface', can play a functional role (Fig. 1d).

Immiscible fluids

Emulsions — droplets of one liquid dispersed in another — have attracted scientific interest ever since Rayleigh in 1879 studied the break up of fluid jets projected on another fluid³¹, with the factors controlling the formation and stability of such droplets established³² by Taylor in 1934. These and other early investigations all used bulk mixtures of immiscible phases to produce large quantities of droplets having a wide range of sizes. Moreover, the practical use of emulsions has long been based on their bulk properties. But with the advent of microfluidics, we can now easily manipulate individual droplets and

precisely control their properties. It has, quite simply, transformed the field.

Dispersion and floating interfaces

A simple microfluidic device for producing and manipulating droplets is the 'T-junction' (Fig. 2a): the T-shaped channel geometry forces two flows of immiscible liquids to merge in such a way that one liquid forms droplets dispersed in the other³³. The droplet-forming phase can be selected by adjusting the hydrophobicity of the device walls at the junction and the relative flow rates of the liquids³⁴. The use of T-junctions in series with alternating surface wettabilities produces monodisperse double emulsions that are useful for encapsulation applications or extractions across the thin layer separating the internal droplets and the continuous phase^{34,35} (Fig. 2b). When reversing the flow direction, T-junctions with differently sized exit channels will passively sort droplets according to size³⁶ or break large droplets into smaller ones with controlled sizes³⁷. Despite its simple design, the T-junction provides precise control over droplet formation (Fig. 2c), making it ideally suited for commercial uses that require parallel, high-throughput predictable droplet creation.

Dispersed droplets may also be created using a microfluidic extension of Rayleigh's approach, with two streams of one liquid flanking a stream of a second immiscible liquid and the combined two-phase flow then forced through a small orifice. The pressure and viscous forces exerted by the outer fluid ultimately force the inner fluid to break into droplets, either in or just downstream of the orifice. The fabrication of a planar microchannel system uses simple soft lithography, making it straightforward to adjust not only flow rates but also the geometry of the microchannel design to ensure selective generation of droplets over a range of different sizes and at different rates^{38,39}. The method is easily adjusted to produce droplets of various compositions (see also Fig. 1a), as demonstrated by the successful synthesis of monodisperse microparticles⁴⁰ and nanoparticles⁴¹ from solutions that allow the droplets to be solidified *in situ* after their formation (by, for example, photopolymerization).

In addition to allowing controlled production of droplets, microfluidic devices also provide an opportunity for precisely manipulating generated droplets. Owing to this combination of capabilities, individual dispersed droplets may serve as floating containers or reaction vessels that can be loaded with different reagents for kinetic measurements⁴²: once a reaction medium has been added and mixed in, the spatial position of the droplet moving continuously along a known path within a microchannel will correlate with reaction time. That is, a given position in the channel will correspond to the same kinetic state, so signals can be collected from several successive droplets and integrated to improve the signal-to-noise ratio, making it possible to monitor even relatively fast reactions with millisecond time resolution or better⁴³. The system is readily extended to studying controlled multi-step reactions by adding new reagents at selected downstream locations⁴⁴ (Fig. 2d). A variation of the method allows efficient screening for optimum protein crystallization conditions by using aqueous droplets in a linear array. Droplets containing different protein solutions alternate with droplets containing salt solutions of different concentrations (see Fig. 2e). Once the array is formed, the difference in osmotic pressure between the alternating static droplets drives the diffusion of water through the oil and thus creates a wide range of different crystallization conditions⁴⁵ while requiring only small quantities of often precious protein material.

Dispersed droplets offer the potential to manipulate or analyse small fluid volumes and thus allow experiments that require only small quantities of reagents (which may be very costly). But the droplet size is so small that solutes will quickly diffuse from the centre to the interface. Depending on composition and affinity, this might result in solutes selectively diffusing out of the droplet or adhering to the interface. If adhesion occurs, the large surface area relative to volume can prove problematic, particularly if the droplet size is decreased so much that adhesion greatly depletes the solute in the droplet interior. In the

case of proteins, adhesion to the droplet interface is often associated with a conformational change, which may become permanent⁴⁶. This tendency to stick to the interface can be prevented by using appropriate surfactants⁴⁷. The effect has also been used to advantage for the fabrication of mechanically stable hollow protein spheres⁴⁸ that might serve as biocompatible 'smart' containers for drug delivery.

At the time of writing, new manipulation methods continue to emerge. For example, microdroplets may be levitated in gas or vacuum using magnetic or acoustic forces⁴⁹, and asymmetric laser heating of the liquid/liquid interface between an aqueous droplet and its surrounding immiscible fluid can induce Marangoni flows to move the droplet⁵⁰. In yet another approach, dispersed droplets are exposed to amphiphilic magnetic nanoparticles that accumulate and align at the droplet interface; the resultant nanoparticle 'coat' then chaperones the liquid droplet in response to an applied external magnetic field⁵¹.

Patterned surfaces and pinned interfaces

A mixture of water and oil in a macroscopic vessel will separate into two phases, with gravitation and the density difference between the fluids ensuring that a horizontal interface separates the oil at the top from the water at the bottom. In micrometre-sized systems, capillary forces can overcome gravitation and be used to create precisely controlled vertical interfaces, or 'virtual walls', between water and air⁵². This requires sufficiently strong capillary forces to 'pin' the water/air interface in position and counteract the action of gravity, which drives water to 'fall' and spread out horizontally. To achieve this, the internal surface of a microchannel is patterned to create hydrophilic and hydrophobic paths. Water molecules will adhere to the hydrophilic channel surface, with surface tension preventing the liquid from invading hydrophobic regions. Aqueous solutions introduced into the patterned microchannel will thus be confined to the hydrophilic regions (see also Fig. 1b), provided the pressure difference across the water/air interface does not exceed a critical value determined by the Young–Laplace equation. The virtual wall between the streams provides a large gas/liquid interface for efficient removal of dissolved gas species such as oxygen from the water stream under continuous operation⁵³. The large surface area provided by virtual walls, and the relatively small volume of the streams to be treated, ensure efficient transport between liquid and gas phases. In this respect, microfluidic devices mimic the alveoli in our lungs, whose large surface area to volume ratio similarly allows rapid exchange of O₂ and CO₂ between air and blood. These systems are not limited to removing dissolved gas from liquid; they could also be used to passively adjust the pH of a buffer solution by exposing it to CO₂ across a virtual wall. Or imagine triggering a chemical reaction within a microfluidic device by introducing a gas-phase species through a virtual wall, or using chemical reactions to generate a gas to be used elsewhere. If airborne analytes are captured into the liquid phase, the system might be used for continuous sensing applications.

It is straightforward to extend the basic idea underlying virtual walls to immiscible liquids flowing side by side (or even on top of each other) in a microchannel. Because the interface between such liquids tends to be unstable owing to differences in liquid properties, patterning the interior microchannel surface to create regions with different wettabilities allows capillary forces to stabilize both vertical⁵⁴ and horizontal⁵⁵ liquid/liquid interfaces. Such 'pinned interfaces' allow for rapid and precise control over the contact time between the two phases, which are typically an organic liquid and an aqueous solution. Moreover, pinned interfaces are produced within seconds, whereas it can take tens of minutes⁵⁶ to establish a stable liquid/liquid interface in a macroscopic system through the action of gravity. These features make microfluidic pinned interfaces attractive for applications such as the study of drug partitioning behaviour⁵⁶ and enzymatic reactions⁵⁷, solvent extraction of metal ions^{58,59}, and the execution of multiphase reactions¹³ and phase-transfer reactions⁶⁰.

Pinned interfaces can also be harnessed more directly. For instance, a stable pinned interface between appropriately chosen aqueous and

organic liquids can serve as the site for an interfacial reaction; if an interfacial polymerization is conducted, the pinned interface is transformed into a real membrane⁵⁴ (Fig. 1b) that is readily functionalized (for instance by immobilizing an enzyme on one of its sides⁶¹). This approach has allowed the formation of a membrane incorporating a peptide crosslinker (N. O. L. Viernes and J. S. Moore, personal communication), so that exposure to a peptide-cleaving enzyme solution leads to a breakdown of the membrane-forming polymer; the resultant fluid leakage then serves as a visual indicator for the presence of the enzyme (Fig. 3). Instead of serving as indicators or sensors, interfacial reactions can also be used to create materials. Particularly when using photopolymerization and suitable surfactant molecules, a wide range of interfaces (including the menisci formed at the solid/liquid/air interface) can be transformed into stable microstructures with unusual shapes, such as microneedles with smooth curved side-walls⁶².

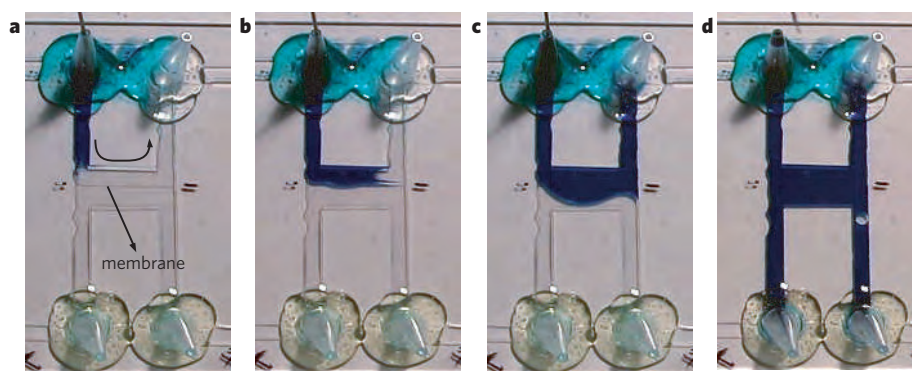


Figure 3 | Pinned interfaces. The use of ‘pinned’ aqueous/organic liquids creates a stable interface where chemistry can occur, for example to create polymer walls. Making use of enzymatic cleavage of peptides, one can create a biosensing or bio-dissolving wall. Here the wall is created through the interfacial reaction of adipoyl chloride and a lysine-terminated peptide creating a thin wall with a known peptide sequence as a crosslink. When exposed to a solution containing appropriate enzymes the peptides are cleaved, the wall becomes more porous and the dyed enzyme-containing solution leaks through the membrane. (Eventually there may be complete dissolution of the wall; not shown.) **a–d**, Sequential images of membrane breakdown. **a, b**, A dyed buffer solution containing chymotrypsin is flowed by capillary action into the top half of the channel. **c**, The solution permeating through the membrane indicates that enzyme cleavage has caused a change in the membrane porosity. **d**, The readout channel is completely filled. Such walls should find use as biosensors with simple visual readouts (as shown) or as intelligent valves that can make process decisions based on changes in the local environment, thereby gating flow to appropriate paths. Courtesy of J. S. Moore, University of Illinois at Urbana-Champaign.

Miscible fluids

The interface between immiscible fluids is easily recognized as the common boundary separating the phases. But if fluids are completely mixed, there is clearly no interface. Still, two miscible fluids, usually liquids, brought into contact will have a boundary between them that disappears as the fluids start mixing. This boundary region can act as a *de facto* dynamic interface that changes with time²² (Fig. 1c) and some of its properties may resemble those of the interface between immiscible fluids^{63,64}. Moreover, if two or more miscible liquids move next to each other under laminar flow conditions, then their diffusive interface can be controlled and used⁶⁵.

Laminar flow and moving interfaces

Laminar flow ensures that mixing between streams in contact with each other occurs only through diffusion. If conditions exist such that the Peclet number is high, mixing will be negligible (see Fig. 1c). At the interface between streams of miscible liquids the contact time is so short that the interface is kinetically stable and remains sharply defined. At lower flow velocity, the liquids are in contact for longer and mix through diffusion: a diffusive interface forms between the fluids, flows and broadens downstream, as the contact time increases.

Laminar flow and diffusion were first put to practical use by Giddings, who used the interface between aqueous streams flowing through microchannels as a ‘virtual membrane’ for protein fractionation^{66–68}. The success of this approach demonstrated that membrane-like performance can be obtained without the potential fouling problems associated with real membranes, and that the effective membrane thickness — the width of the diffusive layer — can be adjusted by simple changes in flow rate. This work, which largely pre-dates what we now regard as microfluidics, used readily available components to create channels of micrometre dimensions in height and millimetre dimensions in width and length. Such simple microchannels suffice for maintaining laminar flow because it is the smallest dimension that largely determines the ratio of inertial to viscous forces (the Reynolds number is a function of channel geometry via hydraulic diameter). That is, in three-dimensional space the scaling of one dimension to the micrometre scale is often sufficient to harness the forces that are dominant at that scale.

Still, the ease with which laminar flow can be realized in modern microfluidic devices allows for particularly effective and precise control over multiple streams of miscible liquids and exploitation of the

interfaces between them. In 1997, microfluidics as we know it today was used to tap into the potential of diffusion and laminar flow^{69,70}, in the shape of the ‘H-filter’. This device merges two separate fluid streams in a central channel and then separates them again into individual channels; the flow regime throughout is laminar. One of the fluids carries particles or solutes of different sizes (the sample stream), while the other is particle-free (the extraction stream). The moment the fluids are in contact, particles start diffusing from the sample to the extraction stream. Diffusivity depends inversely on solute size according to the Stokes–Einstein equation²³, and provided the contact time between the streams is adjusted appropriately, only the smaller solute(s) will enter the extraction stream. Downstream of the central channel, the fluid is split and the extracted solute collected. To achieve fluid splitting without any gross mixing⁶⁹, the two streams need to move with equal velocity and steady flow — conditions that can be challenging to realize experimentally. Moreover, the H-filter requires continually flowing liquids so that the performance of the diffusive interface can be controlled; solutes are therefore extracted at the expense of being diluted.

Like the H filter, the ‘T-sensor’⁷¹ merges two fluid streams into a common channel to create a controlled diffusive interface. One stream typically contains an analyte, the other a tracer compound such as a fluorescent dye or dye-labelled antibody that can interact with the analyte and provide a signal for optical detection. By monitoring the broadening of the interface during the early stages of diffusive mixing, it is then possible to determine diffusion coefficients (from which analyte size can be extracted), analyte concentrations, reaction kinetics and binding affinities⁷². An attractive and useful feature of the T-sensor is that the reagents start to interdiffuse and react the moment the two fluid streams are in contact, so the time available for diffusion and reaction correlates with the distance the fluid travels subsequently. An outside observer will therefore ‘see’ the course of the reaction or diffusion as a still image, and reaction kinetics and diffusion distances can be measured as a function of distance rather than time. This allows the optical signal to be integrated over time to improve sensitivity, making the T-sensor a robust device that is straightforward to implement (in contrast to the H-filter, where the need for stream splitting constitutes a serious complication). At the time of writing, this basic system has been developed for use in molecular mass sensors⁷³, chemical assays⁷⁴, membraneless microfluidic fuel cells^{75,76}, and immunoassays⁷².

The laminar nature of fluid flow through microchannels permits many streams containing not only different substances but also different concentrations of the same substance to flow side by side. As a result, concentration gradients with complex profiles can be generated by feeding a small number of fluid streams with initial concentrations of diffusible substances into a pyramidal microfluidic network^{77,78}. As the streams travel down the network, they are repeatedly split, and some combined with neighbouring streams and allowed to mix by diffusion within a channel before being split and combined again. At the end of the network, many individual fluid streams containing solutions with different concentrations combine in a broad channel that will have a complex concentration profile perpendicular to its flow direction. In contrast to the concentration gradients produced with conventional methods, the profile in this microfluidic system is stable and can be maintained over long periods — characteristics that make the system attractive for studying processes that require gradients, such as chemotaxis^{79,80} and intracellular protein trafficking⁸¹. In contrast to concentration, the local temperature within microfluidic channels can be externally imposed and permanently maintained^{82–85}. This makes it possible to superimpose temperature and concentration gradients in one microfluidic system for high-throughput two-variable experiments.

Laminar flow of multiple liquid streams through microchannels can also serve as a microfabrication tool that is applicable to a broad range of materials, including metals, polymers, inorganic crystals and ceramics^{86,87} (see also Fig. 4a). The chemical composition of the liquids is chosen such that material is deposited onto or etched away from the inner microchannel walls, with the reactions that create the desired structures occurring either between the streams and the contacted channel surface, or at the interface between the streams. For instance, using an etching solution sheathed by an inert liquid results in a narrow trench, the width of which can be adjusted by controlling the relative flow rates of the fluids used. If neighbouring streams carry components that become reactive upon mixing, then etching or material deposition occurs only at locations where the microchannel surface is exposed to the diffusive interface between the streams. Again, adjustment of relative flow rates provides control over the width and location of the structures fabricated.

Laminar flow at high Peclet number, where almost no mixing occurs, is an effective and widely applicable approach for 'sheathing' one fluid with another. It has been used to contain monomers, which

are then photopolymerized 'on the fly' as they exit from the microfluidic device⁸⁸; the method thus allows continuous generation of micrometre-thick fibres and tubes that retain the geometry of the original fluid (Fig. 4b). The process can be used to 'freeze' the diffusive interface, providing insight into directed polymer growth (Fig. 4c). Sheathing is also useful for suppressing non-specific adsorption of analyte to microchannel walls⁸⁹ (a problem that becomes more significant as microfluidic devices are shrunk further).

Zero-flow and secondary interfaces

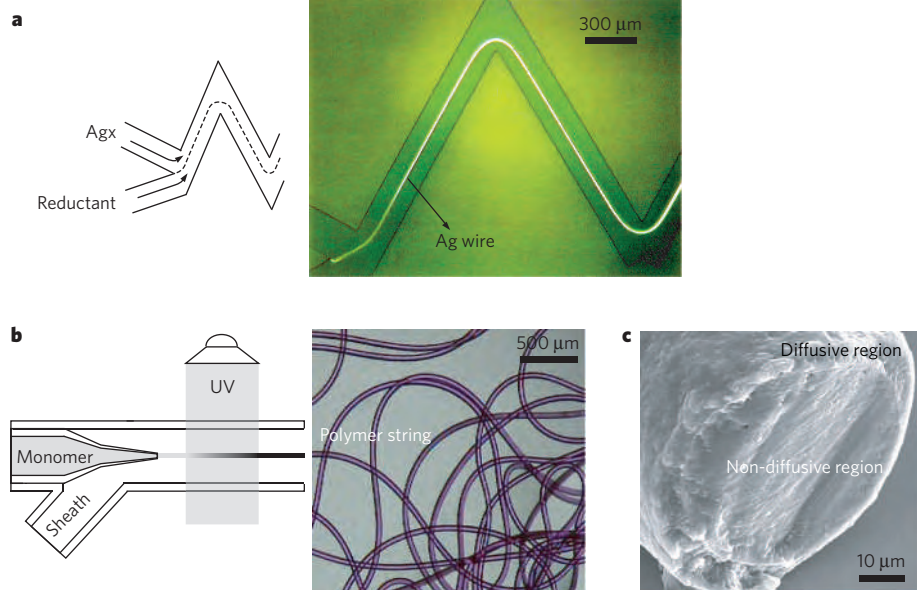
Owing to their small size, microfluidic systems have flow patterns that tend to be dominated by viscous forces; this allows precise control and use of laminar flows. But the dominance of viscous forces makes microfluidic systems also ideally suited for realizing purely diffusive ($Pe = 0$), convection-free environments that are almost impossible to achieve in macroscopic systems (see also Fig. 1d). Only under such conditions will solute released from a source diffuse in all directions with equal probability, its concentration decreasing with increasing distance from the source. The extent of this diffusive layer, which we term a 'secondary interface', depends on the rate of solute release (or production) at the source and on the solute's diffusivity. If microchannel walls are close, solute will accumulate in a predictable way.

A microfluidic chip for screening of protein crystallization conditions¹² takes advantage of such purely diffusive solute transport: compartments are filled with protein solution and precipitant solution, which on opening of a connecting valve mix through diffusion only. Such free interface diffusion is known to make it easier for high-quality crystals to nucleate and grow, but has so far rarely been implemented because of the considerable difficulty of achieving diffusive mixing in large-scale systems.

Flow-free microfluidic systems also offer intriguing opportunities for the study of processes such as cell division and migration, intercellular communication and the emergence of cell polarity during development (where molecular gradients are known to play an important role). For instance, cell proliferation studies on a number of different cell types have revealed that the proliferation characteristics are markedly different when using microchannels instead of traditional mass culture systems^{90,91}. To understand this difference in behaviour, consider the rather different environments experienced by the cells: in the constrained medium within a microchannel, signalling molecules

Figure 4 | Interfacial reactions. Control of the time of contact between two streams in laminar flow is important in these microfabrication processes.

a, Two solutions containing the components of an electroless silver-plating solution flow together in a PDMS microchannel, producing a deposited continuous silver wire at their interface. Image on the left reprinted with permission from ref. 86. Copyright (1999) American Association for the Advancement of Science. Image on the right courtesy of P. J. A. Kenis, University of Illinois at Urbana-Champaign. **b**, Rapid photopolymerization of flowing laminar streams ('on-the-fly' polymerization) allows the continual creation of microscale strings. The smoothness of the interface can be controlled by altering the components in each stream. The addition of multiple sheath flows allows the creation of tubes with controlled size and content⁸⁸. Reproduced by permission of the Royal Society of Chemistry. **c**, Radially directed polymer growth is seen when two miscible fluids are used, one containing a photoinitiator and one without the photoinitiator. The diffusive interface is 'frozen' by photopolymerization allowing high-resolution imaging of the diffusion region between the flows. Courtesy of S. Lee, Dankook University.



produced by a given cell (autocrine signals) or surrounding cells (paracrine signals) can accumulate, whereas such signalling molecules will be diluted or even removed by the convective flows that inevitably arise in mass culture systems or flowing microfluidic systems (see also Fig. 1d). Culturing in microchannels in the absence of flow, where transport is purely by diffusion and the size of the system prevents extensive dilution, appears to increase the sensitivity of proliferating cells to the effects of soluble factors⁹¹. Similar effects may explain why the efficiency of embryo development improves in microchannels under no-flow conditions⁹². So microfluidics should open new opportunities for studying cell signalling, where convection-free culture conditions allow signalling molecules secreted by a cell to form diffusive layers and influence the secreting and surrounding cells. Of course, cells in 'real' living systems are unlikely to experience environments of either laminar flow or no flow at all; still, the ease of creating a range of microenvironmental conditions should allow complementary investigations to characterize and understand cellular processes more fully.

Broadening the range

As we have seen, microfluidics provides us with a powerful way of exploring and exploiting fluid behaviour at a scale where diffusion, viscous drag and surface tension can dominate. Of the applications that are now emerging, we are particularly excited about the unique opportunities for exploring cellular processes. But as the breadth of material and methods presented in this review illustrates, microfluidics can influence a vast range of fields and topics. In fact, whenever we need to use or analyse a fluid, microfluidics could add a new dimension to the task. Cellular autocrine/paracrine signalling mechanisms in development and pathological conditions can now be explored in new ways. Questions of interfacial instabilities and their role in complex systems may become more tractable³³. Basic polymerization dynamics and interfacial reactions can be more carefully examined and turned to better advantage⁸⁷. Basic cellular transport mechanisms might be studied using engineered controlled interfaces to validate proposed models, such as the selective phase model for transport through nuclear pore complexes by hydrophobic exclusion⁹³. It is likely that the future will see a maturing in the way microfluidics are applied: moving beyond the demonstration stage, microfluidics will become an integral tool for formulating and answering these and many other fundamental cross-disciplinary questions. ■

- Vogel, S. *Life in Moving Fluids* (Princeton Univ. Press, Princeton, New Jersey, 1996).
- Purcell, E. M. Life at low Reynolds number. *Am. J. Phys.* **45**, 3–11 (1977).
- Hagen, G. Ueber die Bewegung des Wassers in engen cylindrischen Röhren. *Ann. Phys. Chem.* **46**, 423–442 (1839).
- Poiseuille, J. L. M. Recherches expérimentales sur le mouvement des liquides dans les tubes de très-petits diamètres. *Comptes Rendus* **11**, 961–967 (1841).
- Navier, L. M. H. Mémoire sur les lois du mouvement des fluides. *Mem. Acad. R. Sci.* **6**, 389–440 (1827).
- Stokes, G. G. On the theories of the internal friction of fluids in motion. *Trans. Camb. Phil. Soc.* **8**, 287–319 (1845).
- Taylor, G. I. Dispersion of soluble matter in solvent flowing slowly through a tube. *Proc. R. Soc. Lond. A* **219**, 186–203 (1953).
- Petersen, K. E. Silicon as a mechanical material. *Proc. IEEE* **70**, 420–457 (1982).
- Harrison, D. J. et al. Micromachining a miniaturized capillary electrophoresis-based chemical-analysis system on a chip. *Science* **261**, 895–897 (1993).
- Jacobson, S. C., Hergenroder, R., Koutny, L. B. & Ramsey, J. M. High-speed separations on a microchip. *Anal. Chem.* **66**, 1114–1118 (1994).
- Manz, A., Graber, N. & Widmer, H. M. Miniaturized total chemical-analysis systems: a novel concept for chemical sensing. *Sensors Actuators B* **1**, 244–248 (1990).
- Hansen, C. L., Skordalakes, E., Berger, J. M. & Quake, S. R. A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion. *Proc. Natl Acad. Sci. USA* **99**, 16531–16536 (2002).
- Kobayashi, J. et al. A microfluidic device for conducting gas-liquid-solid hydrogenation reactions. *Science* **304**, 1305–1308 (2004).
- Chaudhry, M. K. & Whitesides, G. M. How to make water run uphill. *Science* **256**, 1539–1541 (1992).
- Zhang, T., Chakraborty, K. & Fair, R. B. *Microelectrofluidic Systems: Modeling and Simulation* (CRC, Boca Raton, 2002).
- Lee, J. & Kim, C. J. Surface-tension-driven microactuation based on continuous electrowetting. *J. Microelectromech. Sys.* **9**, 171–180 (2000).
- Moorthy, J., Khoury, C., Moore, J. S. & Beebe, D. J. Active control of electroosmotic flow in microchannels using light. *Sensors Actuators B* **75**, 223–229 (2001).
- Gascoyne, P. R. C. et al. Dielectrophoresis-based programmable fluidic processors. *Lab Chip* **4**, 299–309 (2004).
- Pollack, M. G., Fair, R. B. & Shenderov, A. D. Electrowetting-based actuation of liquid droplets for microfluidic applications. *Appl. Phys. Lett.* **77**, 1725–1726 (2000).
- Wheeler, A. R., Moon, H., Kim, C. J., Loo, J. A. & Garrell, R. L. Electrowetting-based microfluidics for analysis of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry. *Anal. Chem.* **76**, 4833–4838 (2004).
- Srinivasan, V., Pamula, V. K. & Fair, R. B. An integrated digital microfluidic lab-on-a-chip for clinical diagnostics on human physiological fluids. *Lab Chip* **4**, 310–315 (2004).
- Joseph, D. D. & Renardy, Y. in *Fundamentals of Two-Fluid Dynamics* (eds Antman, S., Marsden, J. E., Sirovich, L. & Wiggins, S.) (Springer, New York, 1993).
- Bird, R. B., Stewart, W. E. & Lightfoot, E. N. *Transport Phenomena* (Wiley, New York, 2001).
- Koschmieder, E. L. in *Bénard Cells and Taylor Vortices* (eds Ablowitz, M. J. et al.) (Cambridge Univ. Press, Cambridge, 1993).
- Gallardo, B. S. et al. Electrochemical principles for active control of liquids on submillimeter scales. *Science* **283**, 57–60 (1999).
- Prins, M. W. J., Welters, W. J. J. & Weekamp, J. W. Fluid control in multichannel structures by electrocapillary pressure. *Science* **291**, 277–280 (2001).
- Adamson, A. W. & Gast, A. P. *Physical Chemistry of Surfaces* (Wiley, New York, 1997).
- Trimmer, W. S. N. Microrobots and micromechanical systems. *Sensors Actuators* **19**, 267–287 (1989).
- Stone, H. A., Stroock, A. D. & Ajdari, A. Engineering flows in small devices: Microfluidics toward a lab-on-a-chip. *Annu. Rev. Fluid Mech.* **36**, 381–411 (2004).
- Squires, T. M. & Quake, S. R. Microfluidics: fluid physics on the nanoliter scale. *Rev. Mod. Phys.* (in the press).
- Rayleigh, L. On the capillary phenomena of jets. *Proc. R. Soc. Lond.* **29**, 71–97 (1879).
- Taylor, G. I. The formation of emulsions in definable fields of flow. *Proc. R. Soc. Lond. A* **146**, 501–523 (1934).
- Thorsen, T., Roberts, R. W., Arnold, F. H. & Quake, S. R. Dynamic pattern formation in a vesicle-generating microfluidic device. *Phys. Rev. Lett.* **86**, 4163–4166 (2001).
- Okushima, S., Nisisako, T., Torii, T. & Higuchi, T. Controlled production of monodisperse double emulsions by two-step droplet breakup in microfluidic devices. *Langmuir* **20**, 9905–9908 (2004).
- Utada, A. S. et al. Monodisperse double emulsions generated from a microcapillary device. *Science* **308**, 537–541 (2005).
- Tan, Y. C., Fisher, J. S., Lee, A. I., Cristini, V. & Lee, A. P. Design of microfluidic channel geometries for the control of droplet volume, chemical concentration, and sorting. *Lab Chip* **4**, 292–298 (2004).
- Link, D. R., Anna, S. L., Weitz, D. A. & Stone, H. A. Geometrically mediated breakup of drops in microfluidic devices. *Phys. Rev. Lett.* **92**, 054503 (2004).
- Anna, S. L., Bontoux, N. & Stone, H. A. Formation of dispersions using 'flow focusing' in microchannels. *Appl. Phys. Lett.* **82**, 364–366 (2003).
- Xu, Q. & Nakajima, M. The generation of highly monodisperse droplets through the breakup of hydrodynamically focused microthread in a microfluidic device. *Appl. Phys. Lett.* **85**, 3726–3728 (2004).
- Jeong, W. J. et al. Continuous fabrication of biocatalyst immobilized microparticles using photopolymerization and immiscible liquids in microfluidic systems. *Langmuir* **21**, 3738–3741 (2005).
- Xu, S. et al. Generation of monodisperse particles by using microfluidics: control over size, shape, and composition. *Angew. Chem. Intl Edn Engl.* **43**, 2–5 (2004).
- Bringer, M. R., Gerdts, C. J., Song, H., Tice, J. D. & Ismagilov, R. F. Microfluidic systems for chemical kinetics that rely on chaotic mixing in droplets. *Phil. Trans. R. Soc. Lond. A* **362**, 1087–1104 (2004).
- Song, H. & Ismagilov, R. F. Millisecond kinetics on a microfluidic chip using nanoliters of reagents. *J. Am. Chem. Soc.* **125**, 14613–14619 (2003).
- Shestopalov, I., Tice, J. D. & Ismagilov, R. F. Multi-step synthesis of nanoparticles performed on millisecond time scale in a microfluidic droplet-based system. *Lab Chip* **4**, 316–321 (2004).
- Zheng, B., Tice, J. D., Roach, L. S. & Ismagilov, R. F. A droplet-based, composite PDMS/glass capillary microfluidic system for evaluating protein crystallization conditions by microbatch and vapor-diffusion methods with on-chip X-ray diffraction. *Angew. Chem. Intl Edn Engl.* **43**, 2508–2511 (2004).
- Dickinson, E. & Matsumura, Y. Proteins at liquid interfaces: role of the molten globule state. *Colloids Surf. B* **3**, 1–17 (1994).
- Roach, L. S., Song, H. & Ismagilov, R. F. Controlling nonspecific protein adsorption in a plug-based microfluidic system by controlling interfacial chemistry using fluorosurfactants. *Anal. Chem.* **77**, 785–796 (2005).
- Lu, G., An, Z. H., Tao, C. & Li, J. B. Microcapsule assembly of human serum albumin at the liquid/liquid interface by the pendent drop technique. *Langmuir* **20**, 8401–8403 (2004).
- Lyuksyutov, I. F., Naugle, D. G. & Rathnayaka, K. D. D. On-chip manipulation of levitated femtodroplets. *Appl. Phys. Lett.* **85**, 1817–1819 (2004).
- Kotz, K. T., Noble, K. A. & Faris, G. W. Optical microfluidics. *Appl. Phys. Lett.* **85**, 2658–2660 (2004).
- Dorvee, J. R., Derfus, A. M., Bhatia, S. N. & Sailor, M. J. Manipulation of liquid droplets using amphiphilic, magnetic one-dimensional photonic crystal chaperones. *Nature Mater.* **3**, 896–899 (2004).
- Zhao, B., Moore, J. S. & Beebe, D. J. Surface-directed liquid flow inside microchannels. *Science* **291**, 1023–1026 (2001).
- Hibara, A. et al. Surface modification method of microchannels for gas-liquid two-phase flow in microchips. *Anal. Chem.* **77**, 943–947 (2005).
- Zhao, B., Viernes, N. O. L., Moore, J. S. & Beebe, D. J. Control and applications of immiscible liquids in microchannels. *J. Am. Chem. Soc.* **124**, 5284–5285 (2002).
- Hibara, A. et al. Stabilization of liquid interface and control of two-phase confluence and separation in glass microchips by utilizing octadecylsilane modification of microchannels. *Anal. Chem.* **74**, 1724–1728 (2002).
- Surmeian, M. et al. Three-layer flow membrane system on a microchip for investigation of molecular transport. *Anal. Chem.* **74**, 2014–2020 (2002).
- Maruyama, T. et al. Enzymatic degradation of p-chlorophenol in a two-phase flow microchannel system. *Lab Chip* **4**, 159–159 (2004).

58. Maruyama, T. *et al.* Intermittent partition walls promote solvent extraction of metal ions in a microfluidic device. *Analyst* **129**, 1008–1013 (2004).
59. Maruyama, T. *et al.* Liquid membrane operations in a microfluidic device for selective separation of metal ions. *Anal. Chem.* **76**, 4495–4500 (2004).
60. Viernes, N. O. L. & Moore, J. S. in *Proc. 7th Int. Conf. Micro Total Analysis Systems* (eds Nothrup, M. A., Jensen, K. F. & Harrison, D. J.) 1041–1044 (Transducers Research Foundation, San Diego/Squaw Valley, 2003).
61. Hisamoto, H. *et al.* Chemicofunctional membrane for integrated chemical processes on a microchip. *Anal. Chem.* **75**, 350–354 (2003).
62. Bauer, J. A., Saif, T. A. & Beebe, D. J. Surface tension driven formation of microstructures. *J. Microelectromech. Syst.* **13**, 553–558 (2004).
63. Garik, P., Hetrick, J., Orr, B., Barkey, D. & Benjacob, E. Interfacial cellular mixing and a conjecture on global deposit morphology. *Phys. Rev. Lett.* **66**, 1606–1609 (1991).
64. Anderson, D. M., McFadden, G. B. & Wheeler, A. A. Diffuse-interface methods in fluid mechanics. *Annu. Rev. Fluid Mech.* **30**, 139–165 (1998).
65. Ismagilov, R. F., Stroock, A. D., Kenis, P. J. A., Whitesides, G. & Stone, H. A. Experimental and theoretical scaling laws for transverse diffusive broadening in two-phase laminar flows in microchannels. *Appl. Phys. Lett.* **76**, 2376–2378 (2000).
66. Giddings, J. C., Yang, F. J. F. & Myers, M. N. Flow field-flow fractionation: versatile new separation method. *Science* **193**, 1244–1245 (1976).
67. Williams, P. S., Levin, S., Lenczycki, T. & Giddings, J. C. Continuous split fractionation based on a diffusion mechanism. *Ind. Eng. Chem. Res.* **31**, 2172–2181 (1992).
68. Giddings, J. C. Field-flow fractionation: analysis of macromolecular, colloidal, and particulate materials. *Science* **260**, 1456–1465 (1993).
69. Brody, J. P. & Yager, P. Diffusion-based extraction in a microfabricated device. *Sensors Actuators A* **58**, 13–18 (1997).
70. Brody, J. P., Yager, P., Goldstein, R. E. & Austin, R. H. Biotechnology at low Reynolds numbers. *Biophys. J.* **71**, 3430–3441 (1996).
71. Weigl, B. H. & Yager, P. Silicon-microfabricated diffusion-based optical chemical sensor. *Sensors Actuators B* **39**, 452–457 (1997); Microfluidics: microfluidic diffusion-based separation and detection. *Science* **283**, 346–347 (1999).
72. Hatch, A. *et al.* A rapid diffusion immunoassay in a T-sensor. *Nature Biotechnol.* **19**, 461–465 (2001).
73. Costin, C. D., McBryde, A. D., McDonnell, M. E. & Synovec, R. E. Theoretical modeling and experimental evaluation of a microscale molecular mass sensor. *Anal. Chem.* **76**, 2725–2733 (2004).
74. Kamholz, A. E., Weigl, B. H., Finlayson, B. A. & Yager, P. Quantitative analysis of molecular interaction in a microfluidic channel: the T-sensor. *Anal. Chem.* **71**, 5340–5347 (1999).
75. Ferrigno, R., Stroock, A. D., Clark, T. D., Mayer, M. & Whitesides, G. M. Membraneless vanadium redox fuel cell using laminar flow. *J. Am. Chem. Soc.* **124**, 12930–12931 (2002).
76. Choban, E. R., Markoski, L. J., Wieckowski, A. & Kenis, P. J. A. Microfluidic fuel cell based on laminar flow. *J. Power Sources* **128**, 54–60 (2004).
77. Jeon, N. L. *et al.* Generation of solution and surface gradients using microfluidic systems. *Langmuir* **16**, 8311–8316 (2000).
78. Dertinger, S. K. W., Chiu, D. T., Jeon, N. L. & Whitesides, G. M. Generation of gradients having complex shapes using microfluidic networks. *Anal. Chem.* **73**, 1240–1246 (2001).
79. Jeon, N. L. *et al.* Neutrophil chemotaxis in linear and complex gradients of interleukin-8 formed in a microfabricated device. *Nature Biotechnol.* **20**, 826–830 (2002).
80. Mao, H. B., Cremer, P. S. & Manson, M. D. A sensitive, versatile microfluidic assay for bacterial chemotaxis. *Proc. Natl Acad. Sci. USA* **100**, 5449–5454 (2003).
81. Sawano, A., Takayama, S., Matsuda, M. & Miyawaki, A. Lateral propagation of EGF signaling after local stimulation is dependent on receptor density. *Dev. Cell* **3**, 245–257 (2002).
82. Mao, H. B., Holden, M. A., You, M. & Cremer, P. S. Reusable platforms for high-throughput on-chip temperature gradient assays. *Anal. Chem.* **74**, 5071–5075 (2002).
83. Ross, D. & Locascio, L. E. Microfluidic temperature gradient focusing. *Anal. Chem.* **74**, 2556–2564 (2002).
84. Pearce, T. M., Wilson, J. A., Oakes, S. G., Chiu, S. Y. & Williams, J. C. Integrated microelectrode array and microfluidics for temperature clamp of sensory neurons in culture. *Lab Chip* **5**, 97–101 (2005).
85. Lucchetta, E. M., Lee, J. H., Fu, L. A., Patel, N. H. & Ismagilov, R. F. Dynamics of *Drosophila* embryonic patterning network perturbed in space and time using microfluidics. *Nature* **434**, 1134–1138 (2005).
86. Kenis, P. J. A., Ismagilov, R. F. & Whitesides, G. M. Microfabrication inside capillaries using multiphase laminar flow patterning. *Science* **285**, 83–85 (1999).
87. Kenis, P. J. A. *et al.* Fabrication inside microchannels using fluid flow. *Acc. Chem. Res.* **33**, 841–847 (2000).
88. Jeong, W. *et al.* Hydrodynamic microfabrication via 'on the fly' photopolymerization of microscale fibers and tubes. *Lab Chip* **4**, 576–580 (2004).
89. Munson, M. S., Hasenbank, M. S., Fu, E. & Yager, P. Suppression of non-specific adsorption using sheath flow. *Lab Chip* **4**, 438–445 (2004).
90. Beebe, D., Wheeler, M., Zeringue, H., Walters, E. & Raty, S. Microfluidic technology for assisted reproduction. *Theriogenology* **57**, 125–135 (2002).
91. Yu, H., Meyvantsson, I., Shkel, I. A. & Beebe, D. Dimension dependent cell behavior in microenvironments. *Lab Chip* (2005).
92. Raty, S. *et al.* Embryonic development in the mouse is enhanced via microchannel culture. *Lab Chip* **4**, 186–190 (2004).
93. Ribbeck, K. & Gorlich, D. The permeability barrier of nuclear pore complexes appears to operate via hydrophobic exclusion. *EMBO J.* **21**, 2664–2671 (2002).

Acknowledgments We thank J. Moorthy and D. Kim for help in preparing the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare that they have no competing financial interests. Correspondence and requests for materials should be addressed to D.J.B. (djbbeebe@wisc.edu).

Polymer-supported membranes as models of the cell surface

Motomu Tanaka^{1,†} & Erich Sackmann¹

Lipid-bilayer membranes supported on solid substrates are widely used as cell-surface models that connect biological and artificial materials. They can be placed either directly on solids or on ultrathin polymer supports that mimic the generic role of the extracellular matrix. The tools of modern genetic engineering and bioorganic chemistry make it possible to couple many types of biomolecule to supported membranes. This results in sophisticated interfaces that can be used to control, organize and study the properties and function of membranes and membrane-associated proteins. Particularly exciting opportunities arise when these systems are coupled with advanced semiconductor technology.

Biological membranes are vital components of all living systems, forming the outer boundary of living cells or internal cell compartments (organelles). They consist largely of a lipid bilayer that imparts a fluid character. Proteins embedded in the bilayer and carbohydrates attached to its surface facilitate communication and transport across the membrane. These features enable membranes to act as important filters: processes (some of which may be incompatible) are confined to the organelles they occur in, and toxic substances are kept out of the cell. But specific nutrients, wastes and metabolites can pass through the membranes of organelles and cells to reach their destination. In addition, many important biological processes are regulated at membrane surfaces, through interactions between peripheral and integral membrane proteins.

The complexity of biological membranes and their interactions with intra- and extracellular networks make direct investigations difficult. For this reason, artificial model membranes have played an important part in unravelling the physical and chemical characteristics of membranes and how these contribute to membrane function. For almost 20 years, phospholipid bilayers deposited onto solid substrates (so-called solid-supported membranes) have been the most commonly used experimental cell-surface model and have allowed us to gain insight into immune reactions and cell adhesion^{1–7}. These model systems are readily prepared by directly depositing lipid monolayers or bilayers onto solid surfaces to yield large areas (of the order of cm²) that maintain excellent mechanical stability without losing their fluid nature^{8–10}. This combination of fluidity and stability on planar surfaces offers distinct advantages over freestanding ‘black’ lipid membranes or spherical lipid vesicles because it makes it possible to carry out experiments and use analytical methods that are difficult or impossible to use with other model systems. For example, methods such as total interference fluorescence^{11,12}, nuclear magnetic resonance (NMR)¹³, Fourier-transform infrared spectroscopy¹⁴, surface plasmon resonance¹⁵ and X-ray and neutron scattering^{16–18} can all be used to probe the structural and dynamic properties of supported membranes.

It is easy to make supported membranes functional by using membrane-associated proteins. This can be achieved in two ways. One method is to spread vesicles containing reconstituted integral proteins,

such as ion channels or membrane-spanning receptors, onto planar substrates. The other method is to prepare supported membranes and incorporate ‘anchor’ molecules and then couple engineered proteins to those anchors. When combined with protein engineering and bioorganic chemistry, these techniques are powerful tools for creating complex experimental cell-surface models that allow us to probe processes that are difficult or even impossible to study otherwise.

Supported membranes containing reconstituted proteins have already provided information about several important biological processes. For example, an early study directly demonstrated that the recognition of antigen-carrying cells by T cells — a crucial feature of the immune response — requires antigens to be associated with the major histocompatibility complex¹¹. Later work⁶ revealed that initiation of the immune response depends on dynamic features of the recognition and interaction process that creates the immunological synapse (the supramolecular contact between the T cell and the antigen-presenting cell). Crucial to the success of this study was the preservation of the lateral fluidity of the lipids within the supported membranes, illustrating that they can serve as artificial surrogate cell surfaces to probe dynamic aspects of membrane function. Moreover, using cells or vesicles in conjunction with supported membranes offers opportunities for examining cell adhesion^{19,20} or the mutual interaction of the proteins that mediate membrane fusion in intracellular vesicle transport and exocytosis²¹.

Solid-supported membranes have some fundamental drawbacks. These arise from the proximity of the artificial membrane and the bare solid surface onto which it is deposited. The membrane–substrate distance is usually not sufficiently large to avoid direct contact between transmembrane proteins incorporated in the membrane and the solid surface. This problem is particularly serious when working with cell-adhesion receptors, whose functional extracellular domains can extend to several tens of nanometres. This problem can be avoided by separating the membrane from the solid substrate using soft polymeric materials that rest on the substrate and support the membrane. This approach significantly reduces the frictional coupling between membrane-incorporated proteins and the solid support, and hence the risk of protein denaturation^{9,21–23}.

¹Department of Physics, Technical University of Munich, D-85748 Garching, Germany. [†]Present address: Institute for Physical Chemistry, University of Heidelberg, D-69120 Heidelberg, Germany.

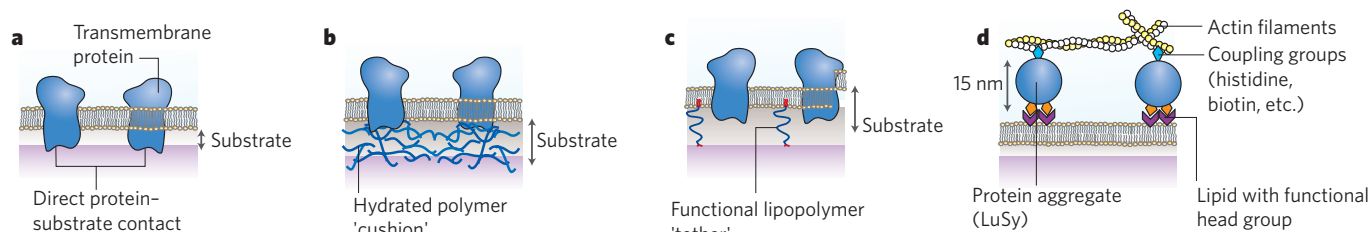


Figure 1 | Supported membranes. The schemes show a solid supported membrane (a), and membranes that are supported using a polymer cushion (b) or lipopolymer tethers (c). Transmembrane proteins are shown in red, membranes in yellow and polymer supports in blue. The thickness of the water reservoir between the membrane and the substrate, distilled H₂O, can be adjusted when using polymer supports. With both types of polymer-supported membrane, the lateral distribution of transmembrane proteins is homogeneous, with proteins tending to exhibit higher diffusivity and improved activity than when incorporated

in solid supported membranes. In the latter, immobile patches caused by the direct protein–solid contact can often be observed. The scheme (d) illustrates actin filaments coupled to a supported membrane using protein aggregates, such as lumazine synthase (LuSy) exposing linker groups (for example, biotin and histidine). The linkers allow attachment of LuSy to membranes containing complementary coupling groups (for example, the chelator complex). Remaining linkers then attach actin filaments, keeping them about 15 nm (the diameter of the capsid) above the membrane surface.

Here, we review different classes of polymer-supported lipid membrane and the various methods for manipulating and patterning them. We also outline the unique opportunities offered by these systems for studying the function of membrane-associated proteins and for practical applications such as protein purification and screening.

Polymer-supported membranes

Supported membranes are readily obtained by depositing phospholipids bilayers onto solid surfaces, with the space between the bilayer and substrate 'bathed' in aqueous solution. In the case of solid-supported membranes, the artificial membrane and its solid support are close together (Fig. 1a). They typically approach each other to within 5–20 Å (refs 13, 17, 24, 25), which leaves a water reservoir that is usually not sufficient to prevent protein subunits from coming into direct contact with the bare substrate. Such direct contact can be avoided by using polymer supports of typically less than 100 nm thickness that 'cushion' or 'tether' the membrane.

Membranes on polymer 'cushions'

When using a polymer to 'cushion' a supported membrane (see Fig. 1b), it should ideally act as a lubricating layer between the membrane and the substrate. This will assist self-healing of local defects in the membrane over macroscopically large substrates (~cm²) and allow the incorporation of large transmembrane proteins without the risk of direct contact between protein subunits and the bare substrate surface.

One of the most important criteria when choosing an appropriate polymer is that the supported membrane must be thermodynamically and mechanically stable, which calls for careful adjustment of the wetting behaviour of the surface-hydrated polymer interface, the surface-membrane interface and the membrane-hydrated polymer interface^{22,26}. The stratified films are only stable if there is complete wetting between the surface and the hydrated polymer, and between the membrane and the hydrated polymer^{27,28}. Moreover, the interaction between the membrane and the surface needs to be repulsive; if it is attractive, the layered system is unstable. This can result in dewetting, giving rise to regions of tight local contact between the membrane and surface (so-called pinning centres). In this regard, polymer cushions mimic the extracellular matrix and the cell-surface glycocalyx, which maintain a relatively high osmotic pressure to keep distinct distances (of typically 10–100 nm) between neighbouring cells and between cells and tissue surfaces. Non-specific contacts caused by van der Waals attraction, which are effective over distances up to about 3 nm, are thus effectively suppressed.

A particularly versatile material for generating polymer cushions is regenerated cellulose²⁹, which can be made into films with flexibly adjustable thickness and wetting properties. The effectiveness of polymer cushions is illustrated with membranes supported on 5-nm-thick

cellulose films and containing cell receptors of the integrin family (human platelet integrin $\alpha_{IIb}\beta_3$). When probing the interaction between these membranes and giant vesicles exposing integrin-specific ligands, the adhesion free energy for the interaction is 3–10-fold higher than the adhesion energy obtained in analogous experiments using solid-supported membranes³⁰ and comparable to the value inferred from the integrin–ligand dissociation constant. Integrins thus seem to fully retain their mobility and native functionality when incorporated in polymer-supported membranes.

Tethers and spacers

In an alternative strategy that separates lipid bilayers from their solid substrates, lipids with macromolecular head groups (so-called lipopolymer tethers) are incorporated into the lipid layer (Fig. 1c). The head groups act as spacers that control the substrate–membrane distance and, in common with polymer cushions, prevent direct contact between transmembrane proteins and their solid substrates. The spacers can be based on a wide range of compounds, including oligo(ethyleneoxide)^{31–33} and poly(ethyleneoxide)³⁴, and oligopeptides with thiol groups³⁵. Living cationic polymerization of poly(2-oxazoline)s yields polymers with precisely controlled monomer numbers, which are attractive as spacers with well-defined lengths^{36,37} for relatively straightforward investigations of the effect of spacer length and lateral spacer density on membrane structure and function³⁶. The ability to flexibly adjust spacer length and lateral spacer density makes it possible to finely tune the membrane–substrate distance and the viscosity of the polymer layer, both of which control the lateral diffusivity and function of transmembrane proteins. This can offer advantages over the use of polymer cushions, where viscosity is completely determined by the material properties of the polymer support itself (see also Box 1).

Another intriguing possibility offered by macromolecular spacers is the attachment of large proteins or protein complexes to supported membranes. If the proteins are attached such that they retain lateral mobility while avoiding non-specific attractive van der Waals contact with the model membrane, then their assembly and interactions can be studied. Particularly attractive macromolecular spacers for this purpose are proteins and their aggregates. For example, the bacterial enzyme lumazine synthase (LUSY) forms a stable 60-mer capsid with a diameter of 15 nm. The monomers can be fused with different linkers (such as histidine and biotin) by expression in bacteria, which enables simple attachment of the assembled capsid to membranes carrying compatible chelator units (see Fig. 1d). Protein complexes can then be coupled to the prepared membrane using the remaining LuSy linkers. An initial demonstration of the system looked at the self-assembly of actin, showing that reversible and controlled binding and unbinding of filaments is possible³⁸. When using supported membranes directly, electrostatic

Box 1 | Diffusion

One of the main benefits of using polymer-supported membranes is that they reduce frictional coupling between proteins and the substrate surface. This makes it possible to easily manipulate functional membrane constituents using tangential electrical fields or temperature gradients, as the drift velocity of proteins is proportional to the passive diffusion coefficient D . Here, measurements of diffusion coefficient D can provide valuable information on the frictional drag acting on proteins.

Macromolecular diffusion in supported membranes differs from that in free membranes (described by Saffman and Delbrück in ref. 86) as a consequence of the frictional tension σ exerted by the solid surface on the diffusing particle, the diffusant. The frictional tension is determined by the viscosity η_c and the thickness h of the hydrated polymer layer, and increases with the velocity v of the diffusant⁸⁵:

$$\sigma = (\eta_c/h)v.$$

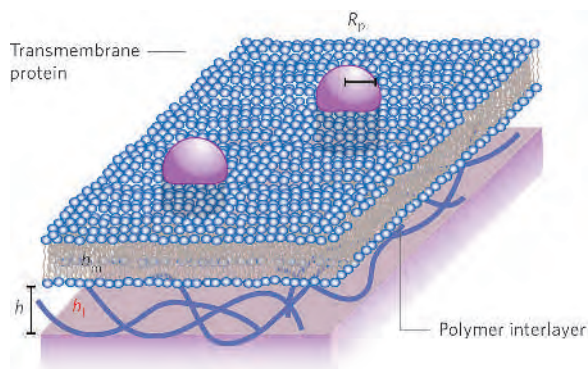
If external frictional forces other than the tension exerted by the surface are negligible, the particle mobility depends only on the frictional tension and the friction within the membrane (with viscosity η_m). In such a case, the diffusion coefficient depends on the radius R_p of the diffusant according to the well known Saffman-Delbrück logarithmic law⁸⁶:

$$D = (k_B T / 8\pi\eta) \ln(R_p^2 \eta_c / \eta_m),$$

where k_B is the Boltzmann constant, T is an absolute temperature, and η is the viscosity of water. For ultrathin polymer cushions ($h = 10$ – 100 nm), the viscosity η_c (and therefore the frictional force σ) is high, and the diffusion coefficient depends very sensitively on the radius of the diffusant:

$$D \approx k_B T h / \eta_c R_p^2.$$

If the thickness of the polymer supporter is known, measurements of passive diffusion coefficients thus enable the inference of either the viscosity of polymer layer η_c or the radius of transmembrane domains R_p (refs 48, 87).



Box 1 Figure | Lateral diffusion of transmembrane proteins in a polymer supported membrane. Here, proteins are assumed to be cylindrical particles, whose transmembrane part has a radius R_p . In the case of ultrathin polymer cushions ($h = 10$ – 100 nm), the passive diffusion constant D can be used as an indicator to calculate either the viscosity of polymer layer η_c or the radius of transmembrane domains R_p .

interactions result in irreversible and ill-defined binding between actin and the supported membrane³⁹.

Supporting native membranes

Nature stringently controls the orientation and the population of transmembrane proteins. It can even dynamically adjust plasma-membrane composition in response to events or stimuli, as demonstrated by the increase in the fraction of a particular protein class (band III proteins) seen upon transfection of a human erythrocyte. Replicating this degree of control using supported membranes is difficult. Transmembrane proteins are usually first stabilized in surfactant micelles and then incorporated into lipid vesicles, which are then used to create supported membranes. Functional assays indicate that

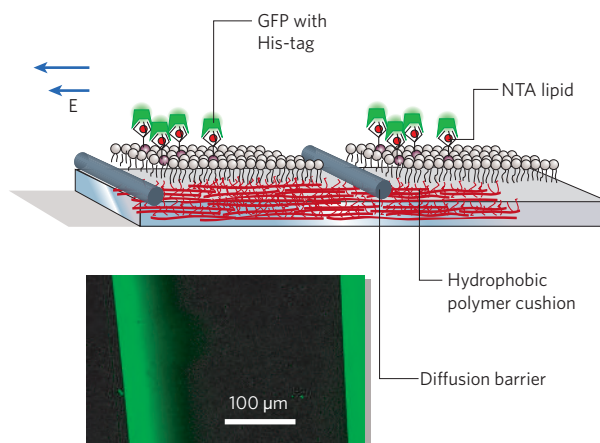


Figure 2 | Electrophoretic accumulation. Electrophoretic accumulation of green fluorescent protein (GFP) with histidine tags (His-EGFP) coupled to a polymer-supported membrane containing 2mol% of chelator lipids. When using hydrophobic cellulose polymer supports, the diffusivities of proteins coupled to membrane lipids are comparable to those of the membrane lipids themselves (1 – $2 \mu\text{m}^2 \text{s}^{-1}$). The accumulation of GFP over distances of several hundreds of μm can be easily achieved by a d.c. electric field of only 10 V cm^{-1} for 30–60 min.

this method can result in directed (orientation-selective) protein incorporation^{21,40}, but many of the surfactants used for protein purification disrupt the membrane so that the incorporated proteins are randomly orientated³⁰. Although the protein:lipid molar ratio of cell membranes is typically 1:500 and can be realized in artificial vesicles with small proteins, it is difficult to go beyond a ratio of 1:5,000 when incorporating large transmembrane proteins.

These problems can be overcome by spreading native cells onto planar substrates, as first demonstrated with human erythrocyte 'ghosts' (cells with their intracellular components removed) and 10-nm-thick, hydrated cellulose cushions. An hour's incubation yielded polymer-supported native membranes that seemed to be defect-free and exposed the cytoplasmic domain⁴¹. The ease and simplicity of the process result from careful initial adjustment of the wetting properties of the system overall, to ensure that membrane spreading is thermodynamically favourable. By contrast, using highly charged polymer films induces dewetting, with cell membranes strongly pinned to the solid substrate⁴¹. Cellulose cushions have also been used for spreading other native membrane extracts such as microsomes from the sarcoplasmic reticulum⁴², and there is no reason to assume that spreading of other types of cell or cell-membrane extract (patches) would not be amenable to this approach as well. We anticipate that when used in conjunction with biochemical assays and suitable detection and monitoring techniques, these systems will allow us to directly probe the intracellular membrane leaflet and processes occurring at or near it.

Manipulating membranes

A characteristic feature of membranes is their fluidity: individual constituents are able to diffuse freely within the membrane plane. As a result, lipids and transmembrane proteins need not always mix uniformly, but are able to cluster to form microdomains. The domains can be relatively static, as seen with lipid rafts and protein clusters, for example, or they can be dynamic, with the accumulation of ligand-receptor complexes mediating adhesion or signalling processes. A distinctive feature of microdomains is that they allow cooperative interactions that enhance overall functionality. For instance, the adhesion of cells onto the walls of venules (small veins) is initiated at several sites within the contact zone, followed by recruitment and accumulation of ligand-receptor pairs to those sites to generate focal-adhesion domains and establish firm adhesion⁴³. Polymer-supported membranes retain much of their fluidity, which makes it possible to create not only static but also dynamic domains.

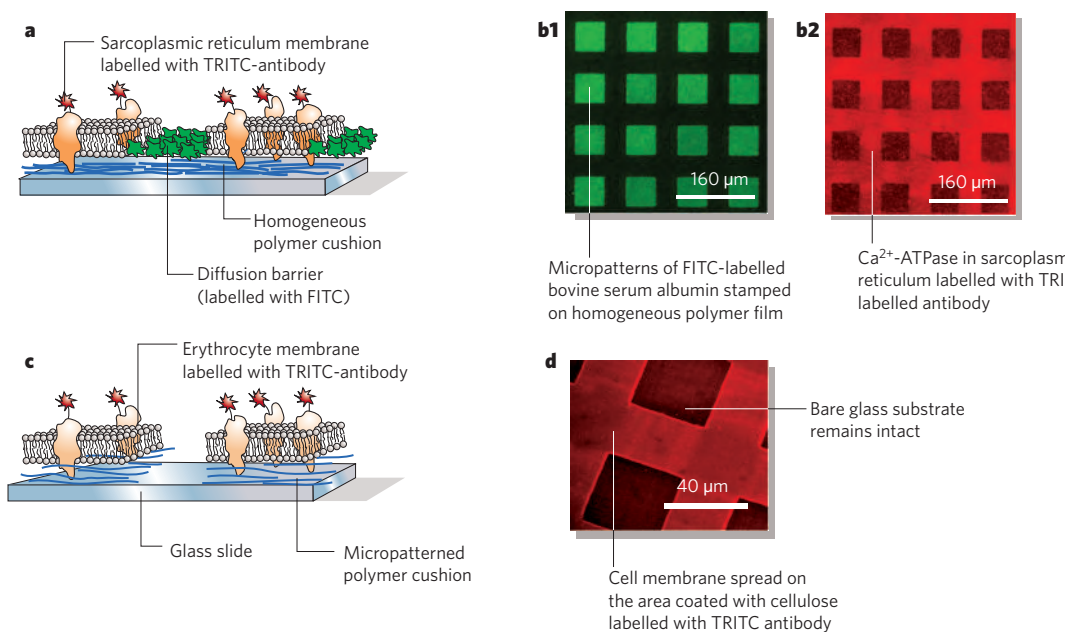


Figure 3 | Membrane patterning. **a**, Sarcoplasmic reticulum membranes confined by diffusion barriers, established by the micro-contact printing of water-soluble protein (bovine serum albumin labelled with FITC, **b1**) on an homogeneous cellulose film. **b**, The cytoplasmic domain of Ca^{2+} -ATPase is visualized with TRITC-labelled antibody (**b2**). **c**, Incubation of human erythrocyte ghosts with lithographically micropatterned cellulose films results in selective spreading of cell membranes on the area coated with cellulose. **d**, Cytoplasmic domain of the proteins in the erythrocyte membrane (band III) is visualized with antibodies conjugated with a fluorescent dye (TRITC).

However, these domains are typically several orders of magnitude larger than those occurring in natural membranes. Rather than mimicking biological processes involving microdomains, these systems are of interest for practical applications such as protein purification and parallel screening.

Electrophoretic accumulation

If a supported membrane contains charged species, these can be accumulated or separated within the membrane environment by applying lateral electric fields. Charged lipids embedded in membranes^{44,45}, the proteins attached to them⁴⁶ and the adsorbed DNA molecules⁴⁷ have all been manipulated in this manner. An intriguing extension involves attaching artificial vesicles with short tethers to lipid headgroups within solid-supported membranes, which allows for transport of synthetic vectors by applying electric fields⁴⁸.

It is the fluid nature of membranes that allows electric fields to work well in altering the distribution of membrane molecules, an effect that might be exploited in electrophoretic separations. When attempting to electrically manipulate lipids, lipid-like proteins and lipid-anchored proteins, crucial issues are the defect density and long-range connectivity of the membranes, with defects being particularly problematic as they can lead to permanent 'traps' that cannot be overcome by electrical fields. Membranes supported on glass have so far proven useful for avoiding these problems, given that many molecules are easily manipulated using these systems. But the approach fails with transmembrane proteins, which are essentially impossible to move in glass-supported bilayers. Extending electrophoretic manipulation to this class of protein is likely to be feasible only if membranes are supported on polymer cushions that avoid contact between membrane protein and substrate.

The proximity between a solid-supported membrane and its support surface is not only problematic when trying to move transmembrane proteins, but also increases epitactic coupling and therefore reduces lateral mobility (see also Box 1)⁴⁹. Polymer cushions can be modified with suitable fluid hydrophobic chains to reduce mechanical coupling between the distal lipid-membrane leaflet and the polymer, thus avoiding electrostatic coupling between the proximal membrane leaflet and the substrate. The effectiveness of this strategy is apparent when comparing covalently grafted alkylsilane monolayer surfaces and suitably designed polymer cushions: lipid monolayers deposited on the latter can have more than 10 times larger lateral diffusivities⁵⁰. With hydrophobic cellulose polymer supports, proteins

coupled to membrane lipids exhibit diffusivities almost independent of protein weight and comparable to the diffusivities of the membrane lipids themselves ($1\text{--}2\ \mu\text{m}^2\ \text{s}^{-1}$). That is, the frictional drag acting on tethered proteins is dominated by the drag acting on the lipid that anchors proteins to the membrane surface. This facilitates easy accumulation of even relatively large proteins that are tethered to membranes and spread over areas stretching over several hundreds of micrometres, using a d.c. electric field of only $10\ \text{V cm}^{-1}$ (see also Fig. 2). For comparison, electric fields typically used in gel electrophoresis are about an order of magnitude stronger. Ongoing membrane separation will of course give rise to inevitable electro-osmotic forces that counteract the electric driving force on the proteins^{46,51}, but enriching the membrane with counterions partly overcomes this problem (J. Hermann, M. Fischer, S. G. Boxer & M.T., unpublished results). To achieve separation of molecules exhibiting only subtle differences in mobility, sophisticated membrane-diffusion-barrier geometries that exploit geometrical brownian ratchet effects⁵² might prove effective.

Patterning membranes

Rather than manipulating the constituents within a membrane, spatial control can also be introduced through patterning to create static and well defined membrane domains or corrals^{10,53}. Micropatterned-domains allow investigations of membrane discrimination⁵⁴, whereas domain arrays offer attractive opportunities for parallel screening of membrane-active analytes, such as antibodies or drugs targeting membrane proteins⁵⁵.

In the case of solid-supported membranes, micrometre-sized patterns can be generated by using a mask to spatially control the photochemical crosslinking of polymerizable lipids⁵⁶ and through micro-contact printing of the membranes themselves⁵⁷. In an alternative and particularly versatile strategy applicable to solid-supported^{58,59} as well as polymer-supported⁴² systems, grid-like diffusion barriers are micro-contact printed using hydrophobic species that attach to the support surface. These barriers then effectively isolate the subsequently deposited membrane into corrals (see also Fig. 3a, b). Patterning is also readily achieved by exploiting wetting contrasts, such as the observation that erythrocyte membranes spread readily on cellulose films but not on glass slides⁴¹. After micropatterned cellulose films with appropriate wetting contrasts have been created⁶⁰, it is straightforward to selectively deposit cell membranes onto the pre-formed patterns⁴² (Fig. 3c, d).

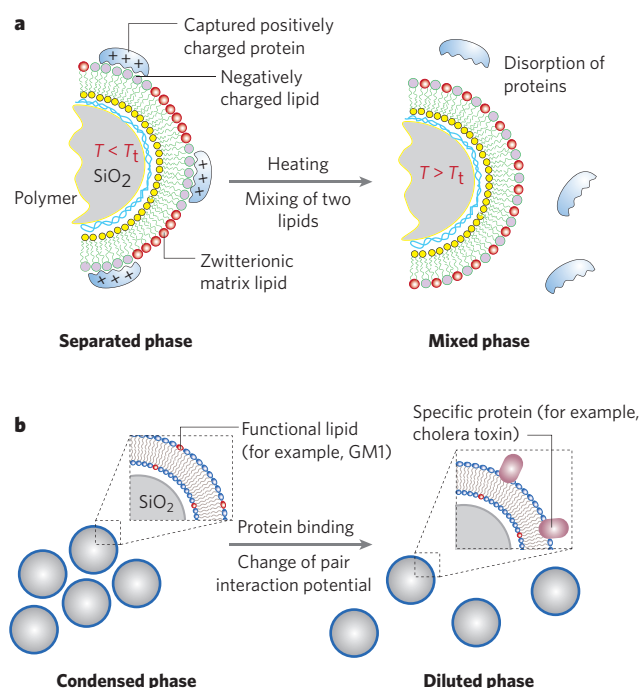


Figure 4 | Membranes on beads. **a**, Principle of 'phase-transition chromatography'. Polymer-coated silica particles are pre-coated with membranes composed of charged and uncharged (zwitterionic) lipids. At low temperature in the gel phase, lateral phase separation creates charged microdomains that electrostatically trap oppositely charged proteins present in an analyte mixture. The trapped proteins can be released by increasing temperatures at which the charged lipid microdomains dissolve. **b**, Use of colloidal phase transition as a label-free detection method for membrane processes. The pair-interaction potential between membrane-coated silica beads is adjusted using an appropriate lipid composition so that the system is poised near its phase transition. Interactions between analytes and membrane components (such as between cholera toxin and ganglioside GM1 in the membrane) change the interaction potential between the beads. Therefore, the system undergoes a phase transition, changing its macroscopic colloidal organization from a condensed to a diluted phase. The overall effect is that the analyte-membrane interaction is amplified, allowing for label-free detection. (See ref. 63 for full details.)

Supported membranes for applications

Transmembrane proteins and membrane-associated proteins are targeted in many infectious diseases. Membrane models play an important part in unravelling the fundamental cellular processes involved and in screening for pathogens or drug candidates. A wide range of membrane systems and detection methods have been developed for this purpose, although polymer-supported membranes are only starting to be used in this context. We will therefore focus here on two general approaches and illustrate their scope with examples that use solid-supported membranes as well as polymer-supported membranes. One of these approaches involves membranes placed on beads or microparticles, which offer some unique advantages for a range of applications. The second approach we will cover aims to electrochemically detect membrane-protein function by using metal and semiconductor electrodes as substrates. In this setup, the electrical properties of the polymer supports play an active role in system design.

Membranes on beads and microparticles

Membranes are easily deposited on silica beads or polymer-coated beads by fusing the beads with lipid vesicles. Bead diameters typically range from 3 to 30 μm , which is comparable to the diameter of cells and about a thousand times larger than the membrane thickness (5–7 nm). The lateral packing and ordering of the lipids on the beads is thus comparable to that of cell membranes.

Some of the advantages of planar-supported membranes for detailed structural and functional investigations are inevitably lost when membranes on beads, but these systems offer a number of other important advantages. For instance, when analytes need to interact with membrane surfaces or membrane-bound agents, the increase in surface area associated with placing membranes on microparticles will significantly increase detection efficiency and speed. Suspensions of membrane-coated microparticles can also be concentrated in any given detection volume; this effect significantly improves the signal of spectroscopic techniques such as NMR¹³, which enables investigations of structural as well as dynamic membrane features.

Like their planar supported counterparts, supported membranes on microparticles are mechanically stable (compared with, say, lipid vesicles) and easily adapted to various functions. Unlike the former, they are readily mixed with reaction or culture media. This attractive combination of features makes it possible to use the particles as reagents. For example, planar supported membranes were used in a study that used membranes containing GPI-anchored neuroligin on silica particles to show that the clustering of β -neuroligin is sufficient to trigger the recruitment of synaptic vesicles⁶¹. The simple fact that membrane-coated beads are mechanically stable and discrete micrometre-sized entities makes it possible to capture them in laser traps⁶². This offers increasingly sophisticated manipulation possibilities (such as the ability to create dynamic three-dimensional arrays) that might lead to unusual applications.

As the approaches mentioned above illustrate, membranes placed directly on microparticles can be used in many different ways. But only if the membranes are supported on polymers is it possible to fully preserve membrane fluidity while presenting a large surface area to analytes. This combination may extend the scope of these systems for high-throughput screening and protein purification. For instance, in one method that makes use of the fluidity of polymer-supported membranes, charged peripheral proteins are trapped and released by using the thermotropic phase transitions exhibited by some lipid mixtures⁶². Here, hydrogel-covered silica particles are coated with lipid membranes containing charged and uncharged lipids. When placed in a chromatography column at low temperature, the charged lipids cluster into domains that can bind oppositely charged target proteins. Heating the column above the phase-transition temperature dissolves the charged lipid domains, which in turn reduces the electrostatic attraction between the membrane and the charged proteins and hence releases them from the beads (Fig. 4a). This method allows the separation of proteins with similar molecular masses but different net charges.

Microparticles can behave not only as mechanically stable discrete entities but also as colloidal particles that exhibit rich colloidal phase behaviour, which can be exploited to monitor molecular interactions⁶³. This requires careful adjustment of membrane composition so that the pair-interaction potential between individual membrane-coated beads poises the system near the transition between a dispersed and condensed phase. As illustrated in Fig. 4b, small membrane perturbations (such as an analyte of antibody binding to a membrane protein) will change the inter-bead potential and induce marked changes in the macroscopic organization of the colloidal phase. This effect enables amplified, label-free detection of cell-membrane processes with high sensitivity⁶⁴.

Native cell membranes can also be deposited on silica microparticles. In fact, the use of microparticles makes it possible to rapidly isolate plasma membranes from intracellular elements^{65–67} as an essential first step towards the analysis of protein composition. This isolation method is particularly attractive because it yields membrane-coated particles with free access to the protoplasmic membrane surfaces for further biochemical studies. In principle, it should be straightforward to use native membranes placed on beads much like their counterparts. For example, they could be used in conjunction with the colloidal phase-transition assay to study protein recognition on the protoplasmic surface of cell membranes.

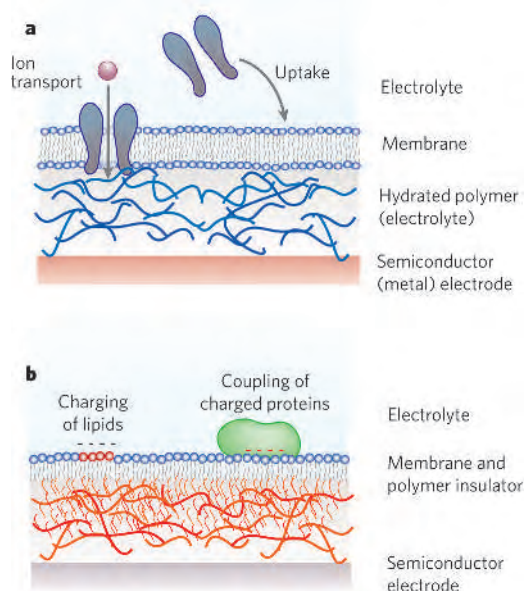


Figure 5 | Electrochemical detection schemes. Electrochemical biosensors using polymer-supported membranes deposited on electrode surfaces. **a**, Using a supported membrane deposited on a hydrated polymer cushion, which electrochemically behaves as an electrolyte, uptake of ion channels and toxins into the supported membrane and their functions can be detected as a change in membrane conductivity. Dynamic Fourier analysis of electrochemical impedance spectra improves the time resolution (to a millisecond order) when monitoring signals from individual ion channels. **b**, The electrolyte-insulator-semiconductor (EIS) structure, which consists of a highly resistant hydrophobic polymer cushion and a supported lipid monolayer. Changes in the surface charge density due to charging of lipids or coupling of charged proteins can be detected quantitatively by monitoring changes in the semiconductor space charge capacitance. Note that when using metal electrodes, the 'total' capacitance cannot be measured with a sensitivity as high as that achieved when using an ITO electrode (one electron per 30 nm^2). The sensitivity achievable with ITO electrodes is sufficient to detect binding of charged peripheral proteins on a single molecular level.

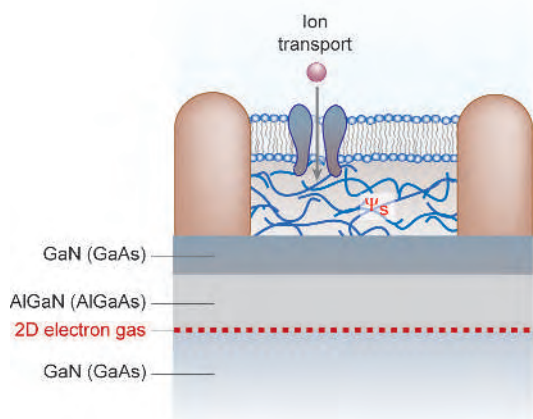


Figure 6 | Membrane-semiconductor systems. The contact between supported membranes and semiconductors can be optimized using suitable polymer supports, whereas patterning techniques ensure the position and area occupied by the supported membranes is matched with the position of semiconductor heterostructure devices (on the basis of GaAs/AlGaAs or GaN/AlGaAs). In such an array, each domain could, in principle, allow for independent local sensing of protein functions. The illustration shows changes in the potential near the surfaces caused by selective ion transport through a channel protein. The potential change can be detected with high sensitivity as a change in the carrier density (that is, the current signal) in a two-dimensional (2D) electron gas layer buried $30\text{--}100 \text{ nm}$ below the semiconductor surface.

Electrochemical detection of protein function

So far, a number of techniques have been used successfully to investigate the function of membrane-associated proteins, often in a quantitative fashion. Fluorescence-based methods in particular have yielded valuable insights owing to their unrivalled sensitivity, which allows detection down to the true single-molecule level. Although there is no doubt that fluorescence detection will play an important role in studies of protein functions in polymer-supported membranes, we focus here on electrochemical methods. These exploit the fact that lipid bilayers possess an intrinsically high electrical resistance and behave essentially as insulators, with the choice of polymer support providing a means to optimize different measurement configurations. In principle, supported membranes deposited on semiconductor electrodes are amenable to two basic measurement strategies: monitoring membrane conductance associated with the transport of ions (conductive sensing; see Fig. 5a), and monitoring changes in membrane surface potential associated with membrane function (capacitive sensing; see Fig. 5b).

In fact, conductive sensing has a long tradition in biology: physiologists widely use the patch-clamp technique⁶⁸ (which involves sealing a fine glass micropipette to a whole cell or small patch of membrane) to monitor the activity of a single or a few ion channels by measuring the associated small (order of picoampere) d.c. currents across the membrane. The method has been crucial for a number of great achievements in our understanding of ion channels, but it is time consuming and limited by the mechanical instability of the clamped membranes. The limited stability of clamped whole cells or membranes is particularly problematic when trying to perform more demanding experiments. The observation that support surfaces containing arrays of cavities with openings of several micrometres significantly improve membrane stability is therefore exciting⁶⁹. Such a set-up has already led to a more comprehensive characterization of structural changes in single-ion channels through simultaneous fluorescence imaging and DC current recordings, revealing that gramicidin dimer formation correlates with channel activity⁷⁰.

Difficulties associated with traditional patch-clamp experiments might also be avoided by placing supported membranes with high mechanical stability onto planar electrode surfaces, provided a high specific membrane resistance (the so-called 'giga-seal', of the order of $1 \text{ M}\Omega \text{ cm}^{-2}$) necessary for electrical detection of the small current signals associated with ion channel activity can be achieved. Although the current passing through a membrane in a d.c. field has not been measured in such systems, a.c.-impedance spectroscopy does enable conductive sensing of the activity of ion channels embedded in supported membranes placed on gold^{32,71,72} and semiconductor^{73–76} electrodes. The complex impedance signal, obtained by measuring the current through the system as a function of frequency, provides information about the electrochemical properties (resistance and capacitance) of the entire system (the membrane, electrode and electrochemical double layers in the electrolyte). In the case of a supported lipid bilayer deposited on a hydrated cellulose cushion (Fig. 5a), the electrochemical property of the polymer film is almost identical to that of the bulk aqueous electrolyte, so the system can electrochemically be treated as an electrolyte-membrane-electrolyte-semiconductor (EMES) multilayer. This makes it possible to determine the electric resistance of the membrane component, which provides quantitative information about selective ion transport via ion channels and carrier proteins incorporated in the membrane. When using indium tin oxide (ITO) electrodes and polymer-supported membranes⁷⁴ with reduced local-defect densities, electric resistances are about 5–50 times higher than can be achieved with solid-supported membranes^{75,77}. As a result, a.c.-impedance spectroscopy with a time resolution of typically 10–60 min is possible; the time resolution can be improved to the millisecond regime by using time-resolved Fourier transform impedance spectroscopy⁷⁸ and a more direct readout of current signals using semiconductor transistors (see also below)^{79–82}.

A hydrophobic cellulose cushion not only provides a fluid environment for a lipid monolayer but also acts as an insulating layer with

a high specific electric resistance of up to $20 \text{ M}\Omega \text{ cm}^2$. A lipid monolayer in contact with an electrolyte and deposited on such a cushion that is sitting on a semiconductor surface thus constitutes an analogue of a metal–insulator–semiconductor (MIS) set-up (Fig. 5b), with the electrolyte acting as the conductor ('metal') and the lipid monolayer and polymer cushion as the insulator. As the film thickness can be adjusted with nanometre accuracy, the potential drop across the polymer insulator can be controlled precisely. The system can thus be tuned so that it is possible to detect the charging and de-charging of the head groups of membrane lipids by monitoring changes in the semiconductor space charge capacitance. When optimized, this simple device should reach a sensitivity of about $1 \times e 30 \text{ nm}^{-2}$ (ref. 50). When considering the size and net charge of typical peripheral proteins, such a sensitivity might well allow the detection of protein binding to the membrane surface on the single-molecule level.

A look to the future

Supported membranes can be patterned and manipulated in many ways to tune their architecture and physical properties for optimal immobilization of peripheral and integral membrane proteins, and to tune the communication between the membrane itself and supporting surfaces. We anticipate that using semiconductors as supports will result in a range of exciting new applications (see also Fig. 6), particularly once it is possible to locally detect signals from individual or small numbers of proteins and enzymes. Imagine depositing membrane micropatterns, doped with different ion channels, on semiconductors that form arrays of field-effect transistors with a sensor area of a few $10 \text{ }\mu\text{m}^2$ each; such a system would allow parallel monitoring of the channel activity in each membrane corral. Combined with microfluidic devices that enable controlled delivery of analytes to individual corrals, this would provide a powerful tool for high-throughput screening.

Crucial to realizing this vision is the availability of highly sensitive, low-dimensional semiconductor structures. With recent advances in band-gap engineering technology, such structures are now increasingly available and ready for use in novel detection schemes. For example, two-dimensional electron gases buried 10–100 nm beneath a semiconductor surface exhibit exquisite sensitivity to small changes in surface-dipole moments and solvent polarities^{83,84}. In fact, first steps towards exploiting the opportunities offered by advanced semiconductor devices have already been taken: the carrier density (that is, conductance) in quantum wires made biologically functional with biotin groups shows a high sensitivity to binding of streptavidin⁸⁵. Moreover, semiconductor field-effect transistors have been used successfully to monitor the activities of neuronal cells^{79,80} and cardiac myocytes⁸². Admittedly, these results serve so far mainly as a proof of principle of what might be possible. But we are excited and confident about the possibilities that can arise from combining supported-membrane technology and semiconductor engineering to develop label-free methods for the detection of protein functions at the molecular level. ■

- Brian, A. A. & McConnell, H. M. Allogeneic stimulation of cytotoxic T cells by supported planar membranes. *Proc. Natl Acad. Sci. USA* **81**, 6159–6163 (1984).
- Chan, P. et al. Influence of receptor lateral mobility on adhesion strengthening between membranes containing LFA-3 and CD2. *J. Cell Biol.* **10**, 245–255 (1991).
- Erb, E.-M., Tangemann, K., Bohrmann, B., Müller, B. & Engel, J. Integrin $\alpha_5\beta_1$ reconstituted into lipid bilayers is nonclustered in its activated state but clusters after fibrinogen binding. *Biochemistry* **36**, 7395–7402 (1997).
- Kloboucek, A., Behrisch, A., Faix, J. & Sackmann, E. Adhesion-induced receptor segregation and adhesion plaque formation: A model membrane study. *Biophys. J.* **77**, 2311–2328 (1999).
- Qi, S. Y., Groves, J. T. & Chakraborty, A. K. Synaptic pattern formation during cellular recognition. *Proc. Natl Acad. Sci. USA* **98**, 6548–6553 (2001).
- Grakoui, A. et al. The immunological synapse: A molecular machine controlling T cell activation. *Science* **285**, 221–227 (1999).
- Tamm, L. K. & McConnell, H. M. Supported phospholipid bilayers. *Biophys. J.* **47**, 105–113 (1985).
- Groves, J. T. & Dustin, M. L. Supported planar bilayers in studies on immune cell adhesion and communication. *J. Immunol. Meth.* **278**, 19–32 (2003).
- Sackmann, E. Supported membranes: Scientific and practical applications. *Science* **271**, 43–48 (1996).
- Groves, J. T. & Boxer, S. G. Micropattern formation in supported lipid membranes. *Acc. Chem. Res.* **35**, 149–157 (2002).
- Watts, T. H., Gaub, H. E. & McConnell, H. M. T-cell-mediated association of peptide antigen and major histocompatibility complex protein detected by energy-transfer in an evanescent wave-field. *Nature* **320**, 179–181 (1986).
- Kalb, E., Frey, S. & Tamm, L. K. Formation of supported planar bilayers by fusion of vesicles to supported phospholipid monolayers. *Biochim. Biophys. Acta* **1103**, 307–316 (1992).
- Bayerl, T. M. & Bloom, M. Physical-properties of single phospholipid-bilayers adsorbed to micro glass-beads — a new vesicular model system studied by H-2-nuclear magnetic-resonance. *Biophys. J.* **58**, 357–362 (1990).
- Tatlian, S. A., Hinterdorfer, P., Baber, G. & Tamm, L. K. Influenza hemagglutinin assumes a tilted conformation during membrane-fusion as determined by attenuated total-reflection FTIR spectroscopy. *EMBO J.* **14**, 5514–5523 (1995).
- Terrettaz, S., Stora, T., Duschl, C. & Vogel, H. Protein-binding to supported lipid-membranes — Investigation of the cholera-toxin ganglioside interaction by simultaneous impedance spectroscopy and surface-plasmon resonance. *Langmuir* **9**, 1361–1369 (1993).
- Kjaer, K., Als-Nielsen, J., Helm, C. A., Laxhuber, L. A. & Mohwald, H. Ordering in lipid monolayers studied by synchrotron X-ray-diffraction and fluorescence microscopy. *Phys. Rev. Lett.* **58**, 2224–2227 (1987).
- Johnson, S. J. et al. Structure of an adsorbed dimyristoylphosphatidylcholine bilayer measured with specular reflection of neutrons. *Biophys. J.* **59**, 289–294 (1991).
- Kalb, E., Engel, J. & Tamm, L. K. Binding of proteins to specific target sites in membranes measured by total internal-reflection fluorescence microscopy. *Biochemistry* **29**, 1607–1613 (1990).
- Bruinsma, R., Behrisch, A. & Sackmann, E. Adhesive switching of membranes: Experiment and theory. *Phys. Rev. E* **61**, 4253–4267 (2000).
- Sackmann, E. & Bruinsma, R. F. Cell adhesion as wetting transition? *Chem. Phys. Chem.* **3**, 262–269 (2002).
- Wagner, M. L. & Tamm, L. K. Reconstituted syntaxin1A/SNAP25 interacts with negatively charged lipids as measured by lateral diffusion in planar supported bilayers. *Biophys. J.* **61**, 266–275 (2001).
- Sackmann, E. & Tanaka, M. Supported membranes on soft polymer cushions: Fabrication, characterization and applications. *Trends Biotechnol.* **18**, 58–64 (2000).
- Knoll, W. et al. Functional tethered lipid bilayers. *Rev. Mol. Biotechnol.* **74**, 137–158 (2000).
- Koenig, B. W. et al. Neutron reflectivity and atomic force microscopy studies of a lipid bilayer in water adsorbed to the surface of a silicon single crystal. *Langmuir* **12**, 1343–1350 (1996).
- Lambacher, A. & Fromherz, P. Fluorescence interference-contrast microscopy on oxidized silicon using a monomolecular dye layer. *Appl. Phys. A* **63**, 207–216 (1996).
- Elender, G. & Sackmann, E. Wetting and dewetting of Si/SiO₂-wafers by free and lipid-monolayer covered aqueous solutions under controlled humidity. *J. Phys. II* **4**, 455–479 (1994).
- Nissen, J., Gritsch, S., Wiegand, G. & Rädler, J. O. Wetting of phospholipid membranes on hydrophilic surfaces — concepts towards self-healing membranes. *Eur. Phys. J. B* **10**, 335–344 (1999).
- Tanaka, M. et al. Wetting and dewetting of extracellular matrix and glycocalyx models. *J. Phys. Cond. Matt.* **17**, S649–S663 (2005).
- Schaub, M., Wenz, G., Wegner, G., Stein, A. & Klemm, D. Ultrathin films of cellulose on silicon wafers. *Adv. Mater.* **5**, 919–922 (1993).
- Goennerwein, S., Tanaka, M., Hu, B., Moroder, L. & Sackmann, E. Functional incorporation of integrins into solid supported membranes on ultrathin films of cellulose: Impact on adhesion. *Biophys. J.* **85**, 646–655 (2003).
- Lang, H., Duschl, C. & Vogel, H. A new class of thiolipid for the attachment of lipid bilayers on gold surfaces. *Langmuir* **10**, 197–210 (1994).
- Cornell, B. A. et al. A biosensor that uses ion-channel switches. *Nature* **387**, 580–583 (1997).
- Schiller, S. M., Naumann, R., Lovejoy, K., Kunz, H. & Knoll, W. Archaea analogue thiolipids for tethered bilayer lipid membranes on ultrasmooth gold surfaces. *Angew. Chem. Int. Ed. Engl.* **42**, 208–211 (2003).
- Wagner, M. L. & Tamm, L. K. Tethered polymer-supported planar lipid bilayers for reconstitution of integral membrane proteins: Silane-polyethyleneglycol-lipid as a cushion and covalent linker. *Biophys. J.* **79**, 1400–1414 (2000).
- Bunjes, N. et al. Thiopeptide-supported lipid layers on solid substrates. *Langmuir* **13**, 6188–6194 (1997).
- Purrucker, O., Förtig, A., Jordan, R. & Tanaka, M. Supported membranes with well-defined polymer tethers — Incorporation of cell receptors. *Chem. Phys. Chem.* **5**, 327–335 (2004).
- Purrucker, O., Förtig, A., Ludke, K., Jordan, R. & Tanaka, M. Confinement of transmembrane receptors in tunable stripe micropatterns. *J. Am. Chem. Soc.* **127**, 1258–1264 (2005).
- Fischer, M., Bacher, A., Haase, I., Tristl, M. & Sackmann, E. Design of biofunctional assemblies on solids through recombinant bacterial protein lumazine synthase. *Chem. Phys. Chem.* **2**, 623–627 (2001).
- Demè, B., Hess, D., Tristl, M., Lee, L.-T. & Sackmann, E. Binding of actin filaments to charged lipid monolayers: Film balance experiments combined with neutron reflectivity. *Eur. Phys. J. E* **2**, 125–136 (2000).
- Salafsky, J., Groves, J. T. & Boxer, S. G. Architecture and function of membrane proteins in planar supported bilayers: A study with photosynthetic reaction centers. *Biochemistry* **35**, 14773–14781 (1996).
- Tanaka, M., Kaufmann, S., Nissen, J. & Hochrein, M. Orientation selective immobilization of human erythrocyte membranes on ultrathin cellulose films. *Phys. Chem. Chem. Phys.* **3**, 4091–4095 (2001).
- Tanaka, M., Wong, A. P., Rehfeldt, F., Tutus, M. & Kaufmann, S. Selective deposition of native cell membranes on biocompatible micro-patterns. *J. Am. Chem. Soc.* **126**, 3257–3260 (2004).
- Springer, T. A. Traffic signals on endothelium for lymphocyte recirculation and leukocyte emigration. *Annu. Rev. Physiol.* **57**, 827–872 (1995).
- Stelzle, M., Mielich, R. & Sackmann, E. Two-dimensional microelectrophoresis in supported lipid bilayers. *Biophys. J.* **63**, 1346–1354 (1992).
- Groves, J. T., Boxer, S. G. & McConnell, H. M. Electric field-induced reorganization of two-component supported bilayer membranes. *Proc. Natl Acad. Sci. USA* **25**, 13390–13395 (1997).

46. Groves, J. T., Wulfiging, C. & Boxer, S. G. Electrical manipulation of glycan phosphatidyl inositol tethered protein in planar supported bilayers. *Biophys. J.* **71**, 2716–2723 (1996).
47. Olson, D. J. *et al.* Electrophoresis of DNA adsorbed to a cationic supported bilayer. *Langmuir* **17**, 7396–7401 (2001).
48. Yoshina-Ishii, C. & Boxer, S. G. Arrays of mobile tethered vesicles on supported lipid bilayers. *J. Am. Chem. Soc.* **125**, 3696–3697 (2003).
49. Merkel, R., Sackmann, E. & Evans, E. Molecular friction and epitactic coupling between monolayers in supported bilayers. *J. Phys. (Paris)* **50**, 1535–1555 (1989).
50. Hillebrandt, H., Tanaka, M. & Sackmann, E. A novel membrane charge sensor: sensitive detection of surface charge at polymer/lipid composite films on indium-tin-oxide electrodes. *J. Phys. Chem. B* **106**, 477–488 (2002).
51. McLaughlin, S. & Poo, M. M. The role of electro-osmosis in the electric-field-induced movement of charged macromolecules on the surfaces of cells. *Biophys. J.* **34**, 85–93 (1981).
52. van Oudenaarden, A. & Boxer, S. G. Brownian ratchets: Molecular separations in lipid bilayers supported on patterned arrays. *Science* **285**, 1046–1048 (1999).
53. Groves, J. T., Ulman, N. & Boxer, S. G. Micropatterning fluid lipid bilayers on solid supports. *Science* **275**, 651–653 (1997).
54. Groves, J. T., Mahal, L. K. & Bertozzi, C. R. Control of cell adhesion and growth with micropatterned supported lipid membranes. *Langmuir* **17**, 5129–5133 (2001).
55. Yang, T., Baryshnikova, O. K., Mao, H., Holden, M. A. & Cremer, P. S. Investigation of bivalent antibody binding on fluid-supported phospholipid bilayers: The effect of hapten density. *J. Am. Chem. Soc.* **125**, 4779–4784 (2003).
56. Morigaki, K., Baumgart, T., Offenhausser, A. & Knoll, W. Patterning solid-supported lipid bilayer membranes by lithographic polymerization of a diacetylene lipid. *Ang. Chem. Inter. Ed.* **40**, 172–174 (2001).
57. Hovis, J. S. & Boxer, S. G. Patterning barriers to lateral diffusion in supported lipid bilayer membranes by blotting and stamping. *Langmuir* **16**, 894–897 (2000).
58. Sapuri, A. R., Baksh, M. M. & Groves, J. T. Electrostatically targeted intermembrane lipid exchange with micropatterned supported membranes. *Langmuir* **19**, 1606–1610 (2003).
59. Kung, L. A., Kam, L., Hovis, J. S. & Boxer, S. G. Patterning hybrid surfaces of proteins and supported lipid bilayers. *Langmuir* **16**, 6773–6776 (2000).
60. Rehfeldt, F. & Tanaka, M. Hydration forces in ultrathin films of cellulose. *Langmuir* **19**, 1467–1473 (2003).
61. Dean, C. *et al.* Neurexin mediates the assembly of presynaptic terminals. *Nature Neurosci.* **6**, 708–716 (2003).
62. Gالنeder, R. *et al.* Microelectrophoresis of a bilayer-coated silica bead in an optical trap: Application to enzymology. *Biophys. J.* **80**, 2298–2309 (2001).
63. Loidl-Stahlhofen, A., Kaufmann, S., Braunschweig, T. & Bayerl, T. M. The thermodynamic control of protein binding to lipid bilayers for protein chromatography. *Nature Biotechnol.* **14**, 999–1002 (1996).
64. Baksh, M. M., Jaros, M. & Groves, J. T. Detection of molecular interactions at membrane surfaces through colloid phase transitions. *Nature* **427**, 139–141 (2004).
65. Jacobson, B. S. & Branton, D. Plasma membrane: rapid isolation and exposure of the cytoplasmic surface by use of positively charged beads. *Science* **195**, 302–304 (1976).
66. Cohen, C. M., Kalish, D. I., Jacobson, B. S. & Branton, D. Membrane isolation on polylysine-coated beads. Plasma membrane from HeLa cells. *J. Cell. Biol.* **75**, 119–134 (1977).
67. Kaufmann, S. & Tanaka, M. Cell adhesion onto highly curved surfaces: One-step immobilization of human erythrocyte membranes on silica beads. *Chem. Phys. Chem.* **4**, 699–704 (2003).
68. Sakmann, B. & Neher, E. *Single-channel Recording* (Plenum, New York, 1985).
69. Fertig, N., Meyer, C., Blick, R. H., Trautmann, C. H. & Behrends, J. C. Microstructured glass chip for ion-channel electrophysiology. *Phys. Rev. E* **64**, 040901 (2001).
70. Borisenko, V. *et al.* Simultaneous optical and electrical recording of single gramicidin channels. *Biophys. J.* **84**, 612–622 (2003).
71. Plant, A. L., Gueguetchkeri, M. & Yap, W. Supported phospholipid/alkanethiol biomimetic membranes — Insulating properties. *Biophys. J.* **67**, 1126–1133 (1994).
72. Steinem, C., Janshoff, A., Ulrich, W.-P., Sieber, M. & Galla, H.-J. Impedance analysis of supported lipid bilayer membranes: a scrutiny of different preparation techniques. *Biochim. Biophys. Acta* **1279**, 169–180 (1996).
73. Stenberg, M., Arwin, H. & Nilsson, A. Silicon-silicon dioxide as an electrode for electrical and ellipsometric measurements of adsorbed organic molecules. *J. Colloid Interface Sci.* **72**, 255–264 (1979).
74. Hillebrandt, H., Wiegand, G., Tanaka, M. & Sackmann, E. High electric resistance polymer/lipid composite films on indium-tin-oxide electrodes. *Langmuir* **15**, 8451–8459 (1999).
75. Gritsch, S., Nollert, P., Jähnig, F. & Sackmann, E. Impedance spectroscopy of porin and gramicidin pores reconstituted into supported lipid bilayers on indium-tin-oxide electrodes. *Langmuir* **14**, 3118–3125 (1998).
76. Purucker, O., Hillebrandt, H., Adlkofer, K. & Tanaka, M. Deposition of highly resistive lipid bilayer on silicon — silicon dioxide electrode and incorporation of gramicidin studied by ac impedance spectroscopy. *Electrochim. Acta* **47**, 791 (2001).
77. Wiegand, G., Arribas-Layton, N., Hillebrandt, H., Sackmann, E. & Wagner, P. Electrical properties of supported lipid bilayer membranes. *J. Phys. Chem. B* **106**, 4245–4254 (2002).
78. Wiegand, W., Neumaier, K. R. & Sackmann, E. Fast impedance spectroscopy: General aspects and performance study for single ion channel measurements. *Rev. Sci. Instrum.* **71**, 2309–2320 (2000).
79. Fromherz, P., Offenhausser, A., Vetter, T. & Weis, J. A neuron-silicon junction — a Retzius cell of the leech on an insulated-gate field-effect transistor. *Science* **252**, 1290–1293 (1991).
80. Straub, B., Meyer, E. & Fromherz, P. Recombinant maxi-K channels on transistor, a prototype of ionic-electronic interfacing. *Nature Biotechnol.* **19**, 121–124 (2001).
81. Steinhoff, G., Purucker, O., Tanaka, M., Stutzmann, M. & Eickhoff, M. AlGa_{1-x}N — A new material system for biosensors. *Adv. Funct. Mater.* **13**, 841–846 (2003).
82. Steinhoff, G. *et al.* Recording of cell action potentials with AlGa_{1-x}N/GaN field-effect transistors. *Appl. Phys. Lett.* **86**, 033901 (2005).
83. Ashkenasy, G., Cahen, D., Cohen, R., Shanzer, A. & Vilan, A. Molecular engineering of semiconductor surfaces and devices. *Acc. Chem. Res.* **35**, 121–128 (2002).
84. Lubner, S. *et al.* Liquid phase sensors based on chemically functionalized GaAs/AlGaAs heterostructures. *Physica E* **21**, 1111–1115 (2004).
85. Cui, Y., Wei, Q., Park, H. & Lieber, C. M. Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. *Science* **293**, 1289–1292 (2001).
86. Saffman, P. G. & Delbrück, M. Brownian motion in biological membranes. *Proc. Natl Acad. Sci. USA* **72**, 3111–3113 (1975).
87. Kühner, M., Tampé, R. & Sackmann, E. Lipid mono- and bilayer supported on polymer films: Composite polymer-lipid films on solid substrates. *Biophys. J.* **67**, 217–226 (1994).

Acknowledgements We thank all our collaborators who contributed to this subject, including S. Kaufmann, O. Purucker, F. Rehfeldt, A. Wong, M. Tutus, J. Hermann, S. Gönnerwein, M. Schneider, K. Adlkofer, H. Hillebrandt, G. Wiegand and S. Gritsch. We thank the groups of G. Wegner, R. Jordan, L. Moroder, M. Fischer, M. Tornow, M. Eickhoff, M. Stutzmann, G. Abstreiter and S. G. Boxer for fruitful collaborations and inspiring discussion. This work was supported through Deutsche Forschungs Gemeinschaft (DFG), National Science Foundation (NSF-MRSEC) and Fonds der Chemischen Industrie.

Author information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to M.T. (mtanaka@ph.tum.de).

Colloidal nanocrystal synthesis and the organic–inorganic interface

Yadong Yin¹ & A. Paul Alivisatos¹

Colloidal nanocrystals are solution-grown, nanometre-sized, inorganic particles that are stabilized by a layer of surfactants attached to their surface. The inorganic cores possess useful properties that are controlled by their composition, size and shape, and the surfactant coating ensures that these structures are easy to fabricate and process further into more complex structures. This combination of features makes colloidal nanocrystals attractive and promising building blocks for advanced materials and devices. Chemists are achieving ever more exquisite control over the composition, size, shape, crystal structure and surface properties of nanocrystals, thus setting the stage for fully exploiting the potential of these remarkable materials.

Colloidal nanocrystals are sometimes referred to as ‘artificial atoms’ because the density of their electronic states — which controls many physical properties — can be widely and easily tuned by adjusting the crystal’s composition, size and shape. The combination of size- and shape-dependent physical properties and ease of fabrication and processing makes nanocrystals promising building blocks for materials with designed functions^{1,2}, for example, as inorganic fluorophores in biomedical assays. But the ability to control the uniformity of the size, shape, composition, crystal structure and surface properties of the nanocrystals is not only of technological interest: access to defined nanoscale structures is essential for uncovering their intrinsic properties unaffected by sample heterogeneity. Rigorous understanding of the properties of individual nanocrystals will enable us to exploit them, making it possible to design and build novel electronic, magnetic and photonic devices and other functional materials based on these nanostructures.

Colloidal nanocrystals have an inorganic core that is stabilized by a layer of surface surfactants. Nanocrystals with a semiconductor as the inorganic material — so-called quantum dots — exhibit size-tunable band gaps and luminescence energies owing to the quantum-size effect³. These colloidal quantum dots are now widely employed as targeted fluorescent labels in biomedical research applications^{4–6}. Compared with the organic fluorophores that were previously used as biological labels, quantum dots are much brighter and do not photobleach. They also provide a readily accessible range of colours. Other applications that could benefit from the combination of low-cost processing and solid-state performance include the use of colloidal quantum dots and rods as alternatives to semiconductor polymers in light emitting diodes⁷, lasers⁸ and solar cells⁹. The scope for these applications has prompted intensive study of the synthesis of these materials to optimize colloidal semiconductor nanocrystal fabrication. As a result, many new concepts for controlling the size, shape and connectivity or coupling of colloidal nanocrystals have been developed first for these materials, but a unified set of synthesis control concepts is now also being applied to other classes of material, such as metals and metal oxides. These materials will extend the range of applications for colloidal nanocrystals to many other areas, including catalysis.

Over the past decade, chemists have come to appreciate that, from

the point of view of synthesis, colloidal inorganic nanocrystals can be thought of as a class of macromolecule, with preparative strategies that are similar in many ways to those employed with artificial organic polymers. For nanocrystals of 1–100 nm diameter, it is possible to define the average and the dispersion of the diameter, as well as the aspect ratio. The degree of precision with which the desired structure is synthesized is similar to that achieved with synthetic polymers, where the preparative methods at our disposal allow us to define the mean number of monomer units in a polymer and the variance of this number, and to build complex topologies through joining or branching of simpler macromolecules. As with artificial polymers, some principles have now emerged that give us the ability to control the size and shape of colloidal nanocrystals. After more than two decades, impressive progress has been made towards the tailored synthesis of colloidal nanocrystals that have well-defined structures. A wide variety can now be successfully produced using a number of methods, such as coprecipitation in aqueous phase, using reverse micelles as templates, hydrothermal/solvothermal synthesis and surfactant-controlled growth in a hot organic solvent^{10,11}.

In this review we outline a set of concepts for controlling the growth of colloidal inorganic nanocrystals in a hot organic surfactant (see also Fig. 1). A general approach to their fabrication in a precisely controlled manner is not yet available, but it is widely accepted that organic surfactants have a key role in determining not only the size but also the shape of the products. Because successful control depends on using organic surfactants to judiciously manipulate the nanocrystal surfaces, we will start with a discussion of the organic–inorganic interface and the possibilities offered by dynamic surface solvation with surfactant molecules. We then introduce the concepts underpinning kinetic control, which allows narrow nanocrystal-size distributions and some control over particle shape. If kinetic control is used in conjunction with selective adhesion effects, it offers even finer control over nanocrystal growth, as illustrated by the strategies used to produce more complex shapes. We conclude this review by outlining the crucial issues that need to be addressed to take colloidal nanocrystal synthesis to the next stage, which will allow the controlled fabrication and processing of ever more complex structures with exciting properties.

¹Department of Chemistry, University of California, Berkeley, and the Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

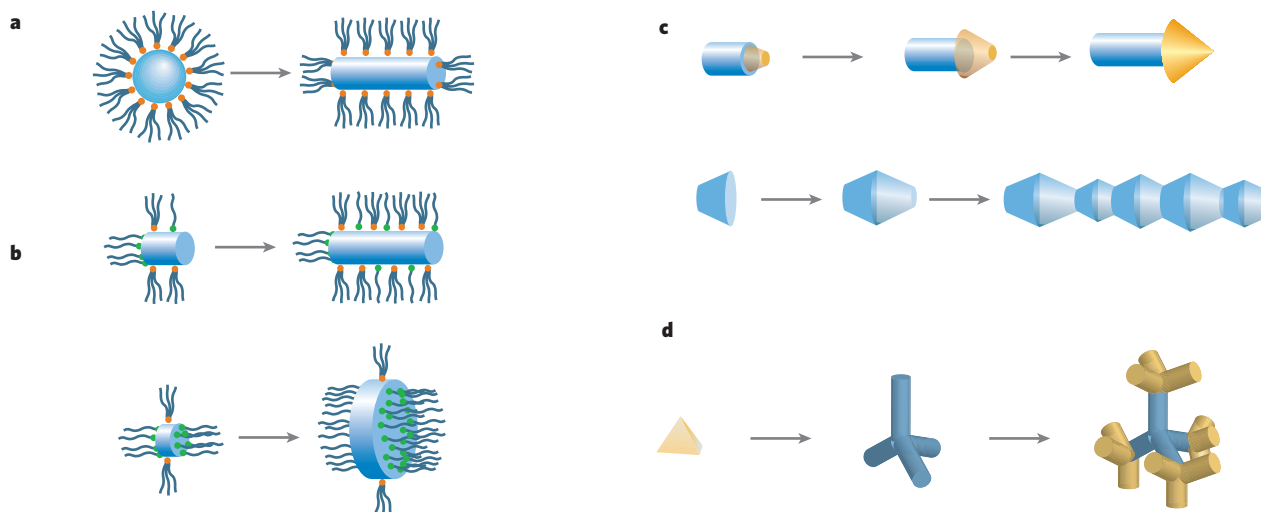


Figure 1 | Shape control of colloidal nanocrystals. **a**, Kinetic shape control at high growth rate. The high-energy facets grow more quickly than low-energy facets in a kinetic regime. **b**, Kinetic shape control through selective adhesion. The introduction of an organic molecule that selectively adheres to a particular crystal facet can be used to slow the growth of that side relative to others, leading to the formation of rod- or disk-shaped nanocrystals. **c**, More intricate shapes result from sequential elimination of a high-energy facet. The persistent growth of an intermediate-energy facet eventually eliminates the initial high-energy facet, forming complex

structures such as an arrow- or zigzag-shaped nanocrystals. **d**, Controlled branching of nanocrystals. The existence of two or more crystal structures in different domains of the same crystal, coupled with the manipulation of surface energy at the nanoscale, can be exploited to produce branched inorganic nanostructures such as tetrapods. Inorganic dendrimers can be further prepared by creating subsequent branch points at the defined locations on the existing nanostructures. The red and green dots in **a** and **b** represent metal coordinating groups with different affinities to nanocrystal facets.

General synthesis scheme

A typical synthesis system for colloidal nanocrystals consists of three components: precursors, organic surfactants and solvents. In some cases, surfactants also serve as solvents. Upon heating a reaction medium to a sufficiently high temperature, the precursors chemically transform into active atomic or molecular species (monomers); these then form nanocrystals whose subsequent growth is greatly affected by the presence of surfactant molecules. The formation of the nanocrystals involves two steps: nucleation of an initial 'seed' and growth. In the nucleation step, precursors decompose or react at a relatively high temperature to form a supersaturation of monomers followed by a burst of nucleation of nanocrystals. These nuclei then grow by incorporating additional monomers still present in the reaction medium.

This thermolysis approach generates nanoparticles that will be crystalline solids only if the constituent atoms can rearrange and anneal during growth. This rearrangement is associated with significant thermal barriers. The cohesive energy per atom, which correlates with the melting temperature of the solid, is therefore a decisive factor in determining optimal conditions for nanocrystal growth. The first step in colloidal nanocrystal synthesis is to choose a temperature for growth that is hot enough to allow rearrangement of atoms and annealing within a growing nanocrystal over the course of the synthesis. A great benefit in this regard is that small crystals require a lower melting temperature. This effect is the subject of one of the most famous and well-documented scaling laws for the properties of solids in the nanometre regime¹². It is driven by the fact that in the nanoscale regime, the liquid phase has lower surface energy than a solid with facets, edges and corners. The effect can be quite significant, leading to a halving of the melting temperature for a solid particle of 2–3 nm diameter relative to that of the corresponding bulk solid.

The large reduction in melting temperature greatly increases the range of inorganic colloidal nanocrystals that can be grown at temperatures where common organic molecules are stable, which is in the range of 200–400 °C. In fact, the desire to extend colloidal nanocrystal synthesis to the widest possible range of materials has focused much interest on growth at these temperatures. Organic surfactant molecules in the growth medium are chosen for their propensity to

adhere to a growing crystal. Because up to half the atoms making up a nanocrystal may be on its surface, this completely alters the growth strategy to the point where the organic–inorganic interface becomes pivotal. Therefore, colloidal nanocrystal growth is strongly related to the field of biomineralization, where complex patterns of biologically organized organic functionalities control the size, shape and spatial arrangement of some inorganic solids^{13,14}. However, the common biominerals produced in this way are strongly ionic solids, with cohesive energies that allow growth at room temperature and in water.

As recognized by Steigerwald^{15,16}, an important step in the generation of colloidal inorganic nanocrystals is the identification of suitable precursor molecules, such as organometallic compounds. The precursors need to rapidly decompose or react at the required growth temperature to yield reactive atomic or molecular species (the monomers), which then cause nanocrystal nucleation and growth. The most famous example of this process is the use of dimethyl cadmium and trialkyl phosphine selenide to yield cadmium selenide (CdSe), where injection of the precursors into a hot solution can yield supersaturation, nucleation and subsequent growth. The most successfully employed precursors have been relatively simple molecules with 'leaving groups' that readily depart leaving behind the desired reactive species. This is somewhat distinct from the precursors employed in chemical vapour deposition (CVD) processes in high vacuum, where volatile precursors react and/or decompose on the substrate surface to produce desired deposit at much higher growth temperatures. Still, the two approaches share many features such as similar basic chemical reactions involved. A literature review of CVD precursors is often a good starting place for finding a new pathway to making a colloidal nanocrystal.

In colloidal solution, the true microscopic mechanism of monomer addition is often still not well understood, owing to the complexity of the growth medium.

Organic–inorganic interface and dynamic solvation

Surfactant-coated nanocrystals, in which an inorganic core is surrounded by a 'monolayer' of organic molecules, hold the potential for the creation of new materials^{17,18}. The possibility of combining the

physical properties of inorganic solids with the low-cost high-volume processing of plastics provides a major impetus for this research¹⁹. Because the organic–inorganic interface present in these systems is the key to the synthesis of far more advanced materials, it is attracting growing interest.

The energy with which surfactant molecules present in the growth medium adhere to the surfaces of growing nanocrystals is one of the most important parameters influencing crystal growth. The adhesion energy needs to be such that it allows dynamic solvation at the growth temperature: the surfactant needs to be able to exchange on and off the growing crystals, so that regions of the nanocrystal surface are transiently accessible for growth, yet entire crystals are, on average, monolayer-protected to block aggregation. The classic paper of Murray, Norris and Bawendi²⁰ introduced this concept for the growth of CdSe nanocrystals in trioctylphosphine oxide (TOPO).

Examples of organic surfactants that dynamically solvate nanocrystals include alkyl phosphine oxides, alkyl phosphonic acids, alkyl phosphines, fatty acids and amines, and some nitrogen-containing aromatics. These molecules all contain metal coordinating groups as well as solvophilic groups. The metal coordinating groups are typically electron-donating to allow coordination to electron-poor metal atoms at the nanocrystal surface. This prevents further growth and aggregation. The other end of the surfactant molecule extends to the solvent and therefore determines the solubility of the nanocrystals; in most cases, it provides the particles with a hydrophobic surface. At the time of writing, there is no generally accepted experimental or theoretical method for determining the adhesion energy of an organic surfactant on a nanocrystal surface, so the choice of surfactant remains empirical. This makes screening techniques borrowed from biochemistry promising tools for discovery of appropriate surfactant systems²¹. In all cases, great care must be taken to examine the purity of the organic surfactants and their thermal stability: in numerous instances small amounts of organic impurities were found to play an important part in the growth kinetics. For example, phosphonic acids were only recognized as essential ingredients for shape control of CdSe and other II–VI nanocrystals²² after their presence in TOPO was finally controlled.

As the temperature decreases, surfactant molecules are less likely to leave the nanocrystal surface. But dynamic solvation can also be achieved at room temperature. For instance, if CdSe nanocrystals coated with TOPO are refluxed in pyridine, the more weakly adherent but more abundant pyridine displaces the TOPO by mass action. Owing to its inherently weaker adhesion energy, pyridine dynamically solvates CdSe, even at room temperature, so that nanocrystals deposited from pyridine and placed in vacuum display ultra-high vacuum (UHV) clean surfaces²³. The introduction of other competing ligands through surfactant exchange makes it possible to further derivatize the nanocrystal surface, allowing the introduction of a wide range of possible chemical functionalities. This strategy provides an additional method for chemical manipulation of the physical nanocrystal properties, which tend to be sensitive to the nature of the surface coating²³. It also provides various means to link the nanocrystals to other surfaces and biomolecules^{24–26}.

The surfactant molecules not only bind to the growing nanocrystal surface, but also form a complex with the reactive monomer species produced upon heating. The stability and diffusion rate of these complexes, as well as the binding strength of the surfactant molecules to the growing nanocrystal surface, are all strongly temperature dependent. Increasing the temperature greatly decreases the stability of the intermediate complexes formed in solution and the binding of the surfactants to the nanocrystal surface, while increasing the diffusion rates of the complexes. This favours the nucleation and growth of the nanocrystals. However, too high a temperature may lead to uncontrolled growth so that it is impossible to exploit subtle kinetic or energetic effects to achieve precise control over the size and size distribution of the nanocrystals. Choosing an appropriate temperature range is one of the key steps in obtaining control over nanocrystal growth.

Kinetic size control

The ability to produce nanocrystals with a relatively narrow size distribution is also a key feature of many modern preparation methods²⁷. To understand the concepts at work here, consider the dependence of growth rate on nanocrystal radius illustrated in Fig. 2. The dependence of the surface energy on size explains the left-hand side of the curve: very small crystals are unstable owing to their large fraction of active surface atoms, as indicated by the negative growth rate. The right-hand side of the curve illustrates that larger crystals with smaller surface-to-volume ratio are stable and grow. The zero-crossing point occurs at the critical size, where nanocrystals neither grow nor shrink. The critical size depends on the monomer concentration, with low monomer concentration favouring a larger critical size. The peak in growth rate versus radius on the right-hand side arises because of a geometric factor: increasing the radius of large crystals requires the incorporation of many more atoms than does increasing the radius of smaller crystals.

These considerations explain why a slow growth rate, which produces equilibrated and nearly round crystals, also yields very broad size distributions. That is, slow growth is associated with low monomer concentrations and a high likelihood that the critical size falls within the distribution of nanocrystal sizes present (Fig. 2). The resultant Ostwald ripening — the shrinking of small crystals while large ones grow — then leads to a broad, skewed size distribution. It is still possible to recover a monodisperse sample by separating out fractions of particles with a narrow distribution from the original broadly distributed sample, using one of several separation techniques. Of these, ‘size-selective precipitation’ is the most generally applicable method^{20,28}. It involves stepwise addition of a poor solvent to a stable solution of nanocrystals to gradually reduce the solvating power and allow for aggregation. Larger nanocrystals, with greater attractive van der Waals or dipolar forces between them, will then precipitate out first. This approach produces nanocrystal fractions with narrow size distributions, but it can be time-consuming, tedious and yields small quantities of the desired material. Further, it only works well with round crystals because the attractive forces between anisotropic nanocrystals depend on multiple parameters²⁹.

A more robust approach to obtaining narrow distributions uses the concept of ‘size-distribution focusing’ (Fig. 2), which is based on the prediction, more than 50 years ago, of Howard Reiss that small crys-

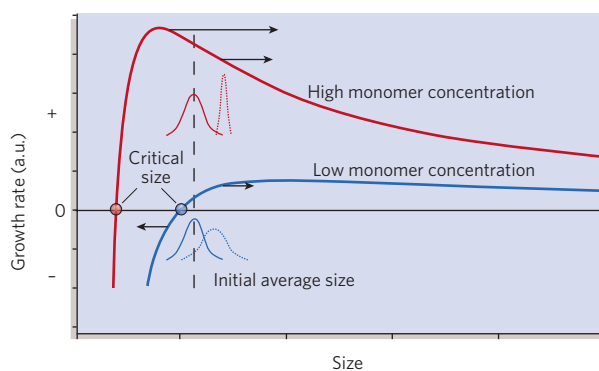


Figure 2 | Size-distribution focusing. The growth process of nanocrystals can occur in two different modes, ‘focusing’ and ‘defocusing’, depending upon the concentration of the monomer present. A critical size exists at any given monomer concentration. At a high monomer concentration, the critical size is small so that all the particles grow. In this situation, smaller particles grow faster than the larger ones, and as a result, the size distribution can be focused down to one that is nearly monodisperse. If the monomer concentration is below a critical threshold, small nanocrystals are depleted as larger ones grow and the size distribution broadens, or defocuses. The preparation of nearly monodisperse spherical particles can be achieved by arresting the reaction while it is still in the focusing regime, with a large concentration of monomer still present. a.u., arbitrary units.

tals will grow more rapidly than larger ones if monomer concentrations are sufficiently high³⁰. Consider the slow growth conditions described above and imagine that the monomer concentration is abruptly increased by a secondary injection of precursor. Immediately after injection, the distribution of nanocrystal sizes present does not change, but the critical size, which depends on monomer concentration, shifts to a smaller value. If this shift is large enough, the entire distribution of sizes will now lie on the falling side of the growth versus radius curve (that is, all nanocrystals are larger than the size for which the growth rate peaks). Therefore, the distribution will spontaneously narrow or 'focus'. The concept of size-distribution focusing has now been clearly demonstrated experimentally³¹. Size focusing is optimal if the monomer concentration is kept such that the average nanocrystal size present is always slightly larger than the critical size. When the monomer concentration is depleted owing to growth, the critical size becomes larger than the average size present, and the distribution broadens as a result of Ostwald ripening. Judicious replenishment of the monomer can thus be an important feature of the synthesis strategy. Focusing has the advantage that it can produce large quantities of crystal with a narrow size distribution, provided that the reaction can be arrested in the appropriate regime (Fig. 3). It is a key first step in kinetic control over nanocrystal synthesis, and we have found that when focusing can be achieved, it opens the door to achieving kinetic shape control.

In general, it is desirable for nucleation to be separated in time from the growth step to obtain relatively monodisperse samples. This means that nucleation must occur on a short time scale. This may be achieved by rapidly injecting suitable precursors into the solvent at high temperature to generate transient supersaturation in monomers and induce a nucleation burst. A rapid and intense nucleation burst will lower the monomer concentration below the nucleation threshold, so monomers remaining in solution will only add to the existing nuclei. In many cases, there is some overlap between the nucleation and growth time scales, so the resultant dispersion in nanocrystal sizes needs to be compensated for with focusing. But in optimal cases, it is possible to remain in the fast growth-focusing regime while remaining below the nucleation limit.

Interestingly, rapid injection of precursor does not always lead to quick nucleation. For example, during the synthesis of iron oxide nanocrystals, the injection of the precursor of iron pentacarbonyl is followed by a long incubation time before a sudden burst of nucleation takes place (M. F. Casula *et al.*, unpublished data). This 'delayed nucleation' is caused by the gradual transformation of iron pentacarbonyl into intermediate species (such as higher nuclearity clusters of carbonyls or metal-surfactant complexes), which then serve as the active 'monomer' species during crystal growth. Because nucleation depends exponentially on the monomer concentration, when the nucleation threshold is surpassed, a brief spurt of nucleation occurs. The nucleation event depletes the monomer, and growth follows with no further nucleation. Delayed nucleation is extremely useful because it removes the need for a rapid (and in many cases irreproducible) initial injection of precursor.

Kinetic shape control

Compared with equilibrium nanocrystals with nearly 'round' shapes, nanocrystals with highly anisotropic shapes have larger surface areas, which renders them metastable, high-energy forms. Formation of the metastable nanocrystals thus requires a kinetic growth regime, whereas equilibrium nanocrystals with low aspect ratios are obtained in the slow growth limit under thermodynamic control. At low growth rate, nearly round nanocrystals are formed, with broad size distribution. At higher growth rate, focusing is observed. When the growth rate is increased just beyond the focusing regime, an astonishing variety of highly anisotropic shapes are obtained, starting with simple rods and disks, but ultimately including shapes like arrows and tetrapods.

The equilibrium shape of inorganic nanocrystals, although faceted, has a low aspect ratio both because this minimizes surface area and

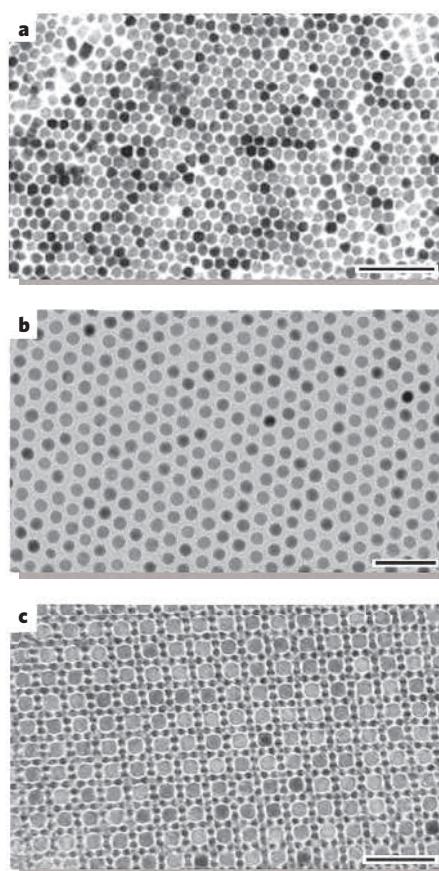


Figure 3 | Monodisperse colloidal nanocrystals synthesized under kinetic size control. **a**, Transmission electron microscopy (TEM) image of CdSe nanocrystals. **b**, TEM image of cobalt nanocrystals. **c**, TEM micrograph of an AB₁₃ superlattice of γ -Fe₂O₃ and PbSe nanocrystals. The precise control on the size distributions of both nanocrystals allows their self-assembly into ordered three-dimensional superlattices. Scale bars, 50 nm. Reprinted from ref. 27.

because the low-energy facets of the crystal are relatively close to each other in energy. However, the growth rate of a crystal facet depends exponentially on the surface energy, so that at high growth rates, in a kinetically controlled growth regime, high-energy facets grow more quickly than low-energy facets (Fig. 1a). The progression from Ostwald ripening to focusing to kinetic shape control was first seen in colloidal CdSe nanocrystals³² and has been subsequently observed in a variety of other systems, such as cobalt^{33,34} and titanium dioxide (TiO₂)³⁵.

The onset of kinetic shape control can be widely adjusted using selective adhesion (Fig. 1b). According to the concept of dynamic solvation, organic surfactants exchange on the nanocrystal's surface during growth. In a faceted crystal, however, the exchange rate on the different facets need not be the same. The introduction of an organic molecule that selectively adheres to a particular crystal facet can be used to effectively lower the energy and slow the growth rate of that facet relative to others (Fig. 4). It is more practical to adjust the relative growth rates than it is to increase the absolute rates to the point where the variations are significant. Selective adhesion effects have not been observed directly during nanocrystal growth, but theoretical studies lend credence to the concept^{36,37}. A possible alternative mechanism, however, involves complexation of the reactive monomer species in solution by organic molecules. This leads to an environment with high chemical potential, and it can be used to adjust relative growth rates³⁸.

In the kinetic growth regime, it is possible to create sequences of events that produce more intricate shapes. A first example is the remarkable phenomenon of sequential elimination of a high-energy facet²² (Fig. 1c). Fast-growing facets will eventually disappear during growth, resulting in a crystal terminated by slower-growing facets. Consider the possibility that the relative growth rates of two different low-index facets differ greatly. In that case, the higher-energy facet will grow so quickly that a second or even third layer of atoms can start to form before a first layer is complete. The possibility exists that there is another facet, intermediate in energy between the low- and high-

energy ones present initially. Such a facet may form transiently during the growth of the high-energy facet. Once such a facet forms, it will persist, replacing the initial high-energy facet. This new intermediate-energy facet will still grow more quickly than the initial slow-growing one, so that the shape will evolve in a complex pattern during growth. This has been used to form arrow-shaped nanocrystals of CdSe²² and zigzag-shaped crystals of TiO₂ (ref. 35; Fig. 5).

A different but related approach to the creation of nanocrystals with complex shapes and connectivity is 'oriented attachment'³⁹. This remarkable process, first described for TiO₂ by Penn and Banfield^{40–42}, involves the coalescence of faceted nanocrystals in such a way as to eliminate two high-energy facets. The detailed mechanism of oriented attachment remains unclear, but the process seems to occur for many materials systems. The most frequent products of oriented attachment are rods and wires. The signature of this mechanism is the observation of final one-dimensional products with the same diameter as the pri-

mary particles, whereas rod lengths are always multiples of the length of the primary nanocrystal. The degree and nature of the attachment process can be manipulated by surfactant control. In fact, specific chemical transformation of the surfactant on high-energy facets may play a crucial role in some of the synthesis strategies reported^{43,44}. Banfield has shown that many defects in natural minerals may have arisen through the oriented attachment process⁴⁵.

Another sequence of events results in controlled branching of colloidal nanocrystals⁴⁶ (Fig. 1d). Branched crystals of zinc oxide were originally discovered in smoke from zinc-smelting plants and have been prepared and studied in CVD systems for several decades⁴⁷. In colloidal systems, ensembles of centrally branched tetrapod nanocrystals can be prepared with a high degree of control over the branch length and diameter^{48,49}. Polytypism⁵⁰, or the existence of two or more crystal structures in different regions of the same crystal, coupled with the manipulation of surface energy at the nanoscale, is exploited to controllably produce the branched inorganic nanostructures. In the case of the II–VI semiconductors such as CdSe, the cubic zincblende ABC stacking of planes is slightly higher in energy but kinetically favoured over the hexagonal wurtzite ABAB stacking. Upon injection of precursors, the high concentration of monomer favours nucleation of a pyramidal seed with a zincblende structure. This seed shares a common crystal facet — the (111) facet — with rod-shaped CdSe with a hexagonal, wurtzite structure. As the monomer concentration drops, the (111) facets of the zincblende core switch to ABAB growth in the [1000] direction of the hexagonal phase. This yields a crystalline inorganic structure of four rods at the tetrahedral angle, a so-called tetrapod (Fig. 6a, b). The presence of a selective adhesion agent that stabilizes the sidewalls of the hexagonal rods relative to the (111) facet of the zincblende phase is a probable reason for the ability to produce these structures with uniformity and control. Most recently, a wide variety of such inorganic dendrimers with branch points at defined locations on rods and tetrapods have been prepared as well⁵¹ (Fig. 6c, d). The subsequent branch points can be created by kinetically driving the reaction again or by nucleation of a second material, such as CdSe on CdTe. These dendritic heterostructures are the most complex structures produced so far in colloidal nanocrystal synthesis.

The outlook

The close topological similarity between inorganic and organic dendrimers helps to emphasize the similarity between organic polymers and inorganic nanocrystals from the point of view of chemistry. Will it ever be possible to create inorganic nanocrystals with the varied and rich compositional and spatial complexity of organic systems? Inorganic nanocrystals can already be created with far more complex shapes and with far greater control over size and shape than had long been thought possible. Colloidal nanocrystals are about the same size as an organic macromolecule, and with organic surfactants on the inorganic nanocrystal surface, they can be manipulated in much the same way as organic polymers⁵². The fact that solid-state materials can be manipulated chemically in such similar ways to polymers has led to a near explosion of work in organic–inorganic colloidal nanocrystals.

But at the time of writing, we are still far from a quantitative description of how organic molecules bind and pack on nanocrystal surfaces. To move beyond the qualitative phenomenology and general framework for kinetic shape control outlined here, detailed knowledge of selective surface-adhesion energies on nanocrystals, including the dependence of adhesion energy on coverage and co-adhesion, is essential. Obtaining such knowledge will, at a minimum, require combined input from experiments on suitable surface science models, appropriate quantum theoretical calculations and detailed structural characterization of colloidal nanocrystal surfaces using new synchrotron-based analysis methods. Improved experimental studies of nanocrystal growth kinetics, including spectroscopic identification of 'monomers' and real-time monitoring of average size and shape, could deliver much-needed further information. In this regard, microfluidics promises exciting opportunities, in that these systems enable

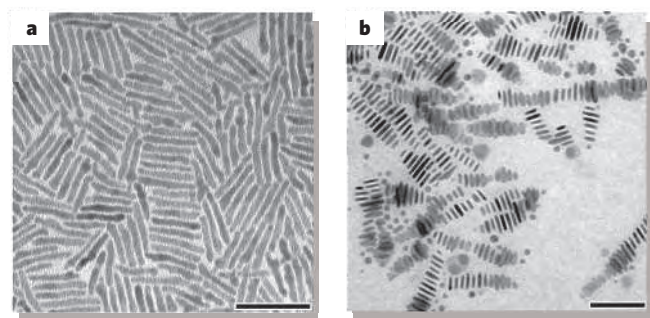


Figure 4 | Anisotropic growth of nanocrystals by kinetic shape control and selective adhesion. **a**, CdSe nanorods (scale bar, 50 nm). Reprinted with permission from ref. 52. **b**, Cobalt nanodisks (scale bar, 100 nm). The organic surfactant molecules selectively adhere to one facet of the nanocrystal, allowing the crystal to grow anisotropically to form a rod or disk.

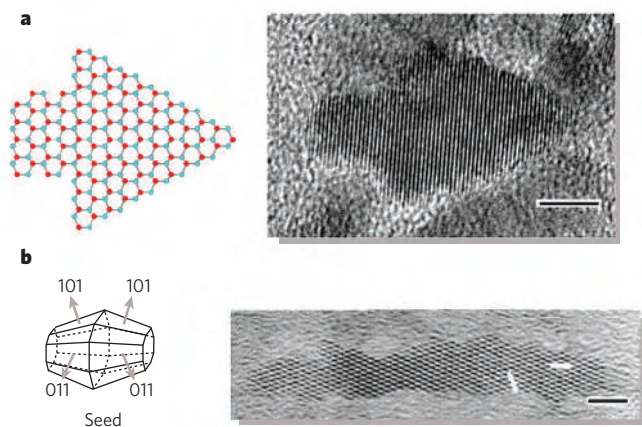


Figure 5 | Nanocrystals with complex shapes prepared by sequential elimination of a high-energy facet. **a**, Two-dimensional representation and a high-resolution TEM image of an arrow-shaped nanocrystal of CdSe. High-resolution TEM characterization shows that each shape of nanocrystal is predominantly wurtzite and that the angled facets of the arrows are the (101) faces. Scale bar, 5 nm. Red and blue dots represent selenium and cadmium atoms, respectively. Reprinted with permission from ref. 22. **b**, Simulated three-dimensional shape and high-resolution TEM analysis of a TiO₂ rod. The long axes of the nanocrystals are parallel to the *c*-axis of the anatase structure, while the nanocrystals are faceted with (101) faces along the short axes. Hexagon shapes (the [010] projection of a truncated octagonal bipyramid) truncated with two (001) and four (101) faces are observed either at the one end or at the centre of the nanocrystals. Scale bar, 3 nm. Reprinted with permission from ref. 35. Copyright (2003) American Chemical Society.

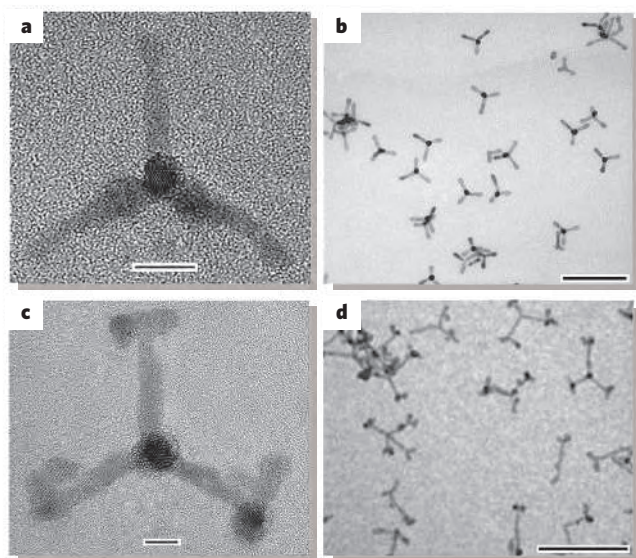


Figure 6 | Controlled branching of colloidal nanocrystals. **a**, High-resolution TEM image of a typical tetrapod-shaped CdSe nanocrystal, looking down the [001] direction of one arm. The nucleus is the zincblende structure, with wurtzite arms growing out of each of the four {111} equivalent faces. Reprinted with permission from ref. 22. **b**, Low-magnification TEM image of CdTe tetrapods. Scale bar, 100 nm. Reprinted from ref. 48. **c**, High-resolution TEM image of a tetrapod that has branches growing out of each arm. There are zincblende layers near the ends of the original arms, and the branches are wurtzite with some stacking faults. Reprinted with permission from ref. 22. **d**, TEM image of branched tetrapods result from nucleation of CdTe zincblende branch points on the end of each arm. Scale bar, 100 nm. Reprinted from ref. 51.

rapid temperature jumps, precise control over concentration as a function of time, and on-chip monitoring and analysis capabilities⁵³.

In the condensed-matter physics community, nanocrystals are commonly referred to as 'artificial atoms', with controlled density of states and designed properties⁵⁴. Staying with this analogy, the ultimate goal of colloidal nanocrystal synthesis is the creation of 'artificial molecules': inorganic nanocrystals with ever more complex yet precisely controlled shapes and compositions, and assemblies of such nanocrystals with carefully positioned interconnections^{27,55–62}. The complexity of such systems would rival that of organic molecules, and they may therefore exhibit a remarkable range of new functionalities. Clearly, exciting opportunities remain in this field, and better understanding and control of the organic–inorganic interface will hold the key to exploring these opportunities and the full potential of colloidal inorganic nanocrystals. ■

- Alivisatos, A. P. Nanocrystals: building blocks for modern materials design. *Endeavour* **21**, 56–60 (1997).
- El-Sayed, M. A. Small is different: shape-, size-, and composition-dependent properties of some colloidal semiconductor nanocrystals. *Acc. Chem. Res.* **37**, 326–333 (2004).
- Alivisatos, A. P. Perspectives on the physical chemistry of semiconductor nanocrystals. *J. Phys. Chem.* **100**, 13226–13239 (1996).
- Bruchez, M., Moronne, M., Gin, P., Weiss, S. & Alivisatos, A. P. Semiconductor nanocrystals as fluorescent biological labels. *Science* **281**, 2013–2016 (1998).
- Michalet, X. *et al.* Quantum dots for live cells, *in vivo* imaging, and diagnostics. *Science* **307**, 538–544 (2005).
- Alivisatos, A. P., Gu, W. & Larabell, C. Quantum dots as cellular probes. *Annu. Rev. Biomed. Eng.* **7**, 55–76 (2005).
- Tessler, N., Medvedev, V., Kazes, M., Kan, S. & Banin, U. Efficient near-infrared polymer nanocrystal light-emitting diodes. *Science* **295**, 1506–1508 (2002).
- Kazes, M., Lewis, D. Y., Ebenstein, Y., Mokari, T. & Banin, U. Lasing from semiconductor quantum rods in a cylindrical microcavity. *Adv. Mater.* **14**, 317–321 (2002).
- Huynh, W. U., Dittmer, J. J. & Alivisatos, A. P. Hybrid nanorod-polymer solar cells. *Science* **298**, 2425–2427 (2002).
- Cushing, B. L., Kolesnichenko, V. L. & O'Connor, C. J. Recent advances in the liquid-phase syntheses of inorganic nanoparticles. *Chem. Rev.* **104**, 3893–3946 (2004).
- Pileni, M. P. The role of soft colloidal templates in controlling the size and shape of inorganic nanocrystals. *Nature Mater.* **2**, 145–150 (2003).
- Buffat, Ph. & Borel, J.-P. Size effect on the melting temperature of gold particles. *Phys. Rev. A* **13**, 2287–2298 (1976).

- Mann, S. Molecular recognition in biomineralization. *Nature* **332**, 119–124 (1988).
- Bianconi, P. A., Lin, J. & Strzelecki, A. R. Crystallization of an inorganic phase controlled by a polymer matrix. *Nature* **349**, 315–317 (1991).
- Stuczynski, S. M., Brennan, J. G. & Steigerwald, M. L. Formation of metal–chalcogen bonds by the reaction of metal–alkyls with silyl chalcogenides. *Inorg. Chem.* **28**, 4431–4432 (1989).
- Steigerwald, M. L. Clusters as small solids. *Polyhedron* **13**, 1245–1252 (1994).
- Nirmal, M. & Brus, L. Luminescence photophysics in semiconductor nanocrystals. *Acc. Chem. Res.* **32**, 407–414 (1999).
- Alivisatos, A. P. Semiconductor clusters, nanocrystals, and quantum dots. *Science* **271**, 933–937 (1996).
- Steckel, J. S., Coe-sullivan, S., Bulovic, V. & Bawendi, M. G. 1.3 μm to 1.55 μm tunable electroluminescence from PbSe quantum dots embedded within an organic device. *Adv. Mater.* **15**, 1862–1866 (2003).
- Murray, C. B., Norris, D. J. & Bawendi, M. G. Synthesis and characterization of nearly monodisperse CdE (E = sulfur, selenium, tellurium) semiconductor nanocrystallites. *J. Am. Chem. Soc.* **115**, 8706–8715 (1993).
- Whaley, S. R., English, D. S., Hu, E. L., Barbara, P. F. & Belcher, A. M. Selection of peptides with semiconductor binding specificity for directed nanocrystal assembly. *Nature* **405**, 665–668 (2000).
- Manna, L., Scher, E. C. & Alivisatos, A. P. Synthesis of soluble and processable rod-, arrow-, teardrop-, and tetrapod-shaped CdSe nanocrystals. *J. Am. Chem. Soc.* **122**, 12700–12706 (2000).
- Katari, J. E. B., Colvin, V. L. & Alivisatos, A. P. X-ray photoelectron spectroscopy of CdSe nanocrystals with applications to studies of the nanocrystal surface. *J. Phys. Chem.* **98**, 4109–4117 (1994).
- Kuno, M., Lee, J. K., Dabbousi, B. O., Mikulec, F. V. & Bawendi, M. G. The band edge luminescence of surface modified CdSe nanocrystallites: probing the luminescing state. *J. Chem. Phys.* **106**, 9869–9882 (1997).
- Klein, D. L., Roth, R., Lim, A. K. L., Alivisatos, A. P. & McEuen, P. L. A single-electron transistor made from a cadmium selenide nanocrystal. *Nature* **389**, 699–701 (1997).
- Chan, W. C. & Nie, S. Quantum dot bioconjugates for ultrasensitive nonisotopic detection. *Science* **281**, 2016–2018 (1998).
- Redl, F. X., Cho, K.-S., Murray, C. B. & O'Brien, S. Three-dimensional binary superlattices of magnetic nanocrystals and semiconductor quantum dots. *Nature* **423**, 968–971 (2003).
- Vossmeier, T. *et al.* CdS nanoclusters: synthesis, characterization, size dependent oscillator strength, temperature shift of the excitonic transition energy, and reversible absorbance shift. *J. Phys. Chem.* **98**, 7665–7673 (1994).
- Murray, C. B., Sun, S., Doyle, H. & Betley, T. Monodisperse 3d transition-metal (Co, Ni, Fe) nanoparticles and their assembly into nanoparticle superlattices. *Mater. Res. Soc. Bull.* **26**, 985–991 (2001).
- Reiss, H. The growth of uniform colloidal dispersions. *J. Chem. Phys.* **19**, 482–487 (1951).
- Peng, X., Wickham, J. & Alivisatos, A. P. Kinetics of II–VI and III–V colloidal semiconductor nanocrystal growth: focusing of size distributions. *J. Am. Chem. Soc.* **120**, 5343–5344 (1998).
- Peng, X. *et al.* Shape control of CdSe nanocrystals. *Nature* **404**, 59–61 (2000).
- Puntes, V. F., Krishnan, K. M. & Alivisatos, A. P. Colloidal nanocrystal shape and size control: the case of cobalt. *Science* **291**, 2115–2117 (2001).
- Puntes, V. F., Zanchet, D., Erdonmez, C. K. & Alivisatos, A. P. Synthesis of hcp-Co nanodisks. *J. Am. Chem. Soc.* **124**, 12874–12880 (2002).
- Jun, Y.-W. *et al.* Surfactant-assisted elimination of a high energy facet as a means of controlling the shapes of TiO₂ nanocrystals. *J. Am. Chem. Soc.* **125**, 15981–15985 (2003).
- Puzder, A. *et al.* The effect of organic ligand binding on the growth of CdSe nanoparticles probed by ab initio calculations. *Nano Lett.* **4**, 2361–2365 (2004).
- Manna, L., Wang, L. W., Cingolani, R. & Alivisatos, A. P. First-principles modeling of unpassivated and surfactant-passivated bulk facets of wurtzite cdse: a model system for studying the anisotropic growth of CdSe nanocrystals. *J. Phys. Chem. B* **109**, 6183–6192 (2005).
- Yu, W. W., Wang, Y. A. & Peng, X. Formation and stability of size-, shape-, and structure-controlled CdTe nanocrystals: ligand effects on monomers and nanocrystals. *Chem. Mater.* **15**, 4300–4308 (2003).
- Alivisatos, A. P. Naturally aligned nanocrystals. *Science* **289**, 736–737 (2000).
- Penn, R. L. & Banfield, J. F. Morphology development and crystal growth in nanocrystalline aggregates under hydrothermal conditions: insights from titania. *Geochim. Cosmochim. Acta* **63**, 1549–1557 (1999).
- Penn, R. L. & Banfield, J. F. Oriented attachment and growth, twinning, polytypism, and formation of metastable phases: insights from nanocrystalline TiO₂. *Am. Mineral.* **83**, 1077–1082 (1998).
- Pacholski, C., Kornowski, A. & Weller, H. Self-assembly of ZnO: from nanodots to nanorods. *Angew. Chem. Int. Ed.* **41**, 1188–1191 (2002).
- Yu, J. H. *et al.* Synthesis of quantum-sized cubic ZnS nanorods by the oriented attachment mechanism. *J. Am. Chem. Soc.* **127**, 5662–5670 (2005).
- Adachi, M. *et al.* Highly efficient dye-sensitized solar cells with a titania thin-film electrode composed of a network structure of single-crystal-like TiO₂ nanowires made by the "oriented attachment" mechanism. *J. Am. Chem. Soc.* **126**, 14943–14949 (2004).
- Banfield, J. F. & Penn, R. L. Imperfect oriented attachment: dislocation generation in defect-free nanocrystals. *Science* **281**, 969–971 (1998).
- Wang, D. & Lieber, C. M. Nanocrystals branch out. *Nature Mater.* **2**, 355–356 (2003).
- Yan, H., He, R., Pham, J. & Yang, P. Morphogenesis of one-dimensional ZnO nano- and microcrystals. *Adv. Mater.* **15**, 402–405 (2003).
- Manna, L., Milliron, D. J., Meisel, A., Scher, E. C. & Alivisatos, A. P. Controlled growth of tetrapod branched inorganic nanocrystals. *Nature Mater.* **2**, 382–385 (2003).
- Peng, X. Mechanisms for the shape-control and shape-evolution of colloidal semiconductor nanocrystals. *Adv. Mater.* **15**, 459–463 (2003).
- Yeh, C. Y., Lu, Z. W., Froyen, S. & Zunger, A. Zinc-blende-wurtzite polytypism in semiconductors. *Phys. Rev. B* **46**, 10086–10097 (1992).
- Milliron, D. J. *et al.* Colloidal nanocrystal heterostructures with linear and branched topology. *Nature* **430**, 190–195 (2004).
- Li, L.-S., Walda, J., Manna, L. & Alivisatos, A. P. Semiconductor nanorod liquid crystals. *Nano Lett.* **2**, 558–560 (2002).

53. Chan, E. M., Mathies, R. A. & Alivisatos, A. P. Size-controlled growth of CdSe nanocrystals in microfluidic reactors. *Nano Lett.* **3**, 199–201 (2003).
54. Collier, C. P., Vossmeier, T. & Heath, J. R. Nanocrystal superlattices. *Annu. Rev. Phys. Chem.* **49**, 371–404 (1998).
55. Sun, S., Murray, C. B., Weller, D., Folks, L. & Moser, A. Monodisperse FePt nanoparticles and ferromagnetic FePt nanocrystal superlattices. *Science* **287**, 1989–1992 (2000).
56. Mirkin, C. A., Letsinger, R. L., Mucic, R. C. & Storhoff, J. J. A DNA-based method for rationally assembling nanoparticles into macroscopic materials. *Nature* **382**, 607–609 (1996).
57. Alivisatos, A. P. *et al.* Organization of 'nanocrystal molecules' using DNA. *Nature* **382**, 609–611 (1996).
58. Fu, A. *et al.* Discrete nanostructures of quantum dots/Au with DNA. *J. Am. Chem. Soc.* **126**, 10832–10833 (2004).
59. Yu, H. *et al.* Dumbbell-like bifunctional Au-Fe₃O₄ nanoparticles. *Nano Lett.* **5**, 379–382 (2005).
60. Gu, H., Zheng, R., Zhang, X. & Xu, B. Facile one-pot synthesis of bifunctional heterodimers of nanoparticles: a conjugate of quantum dot and magnetic nanoparticles. *J. Am. Chem. Soc.* **126**, 5664–5665 (2004).
61. Mokari, T., Rothenberg, E., Popov, I., Costi, R. & Banin, U. Selective growth of metal tips onto semiconductor quantum rods and tetrapods. *Science* **304**, 1787–1790 (2004).
62. Kudara, S. *et al.* Selective growth of PbSe on one or both tips of colloidal semiconductor nanorods. *Nano. Lett.* **5**, 445–449 (2005).

Acknowledgements The authors acknowledge support from the US Department of Energy through the Molecular Foundry at the Lawrence Berkeley National Laboratory.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.A. (alivis@berkeley.edu).

Engineering atomic and molecular nanostructures at surfaces

Johannes V. Barth^{1,2}, Giovanni Costantini³ & Klaus Kern^{1,3}

The fabrication methods of the microelectronics industry have been refined to produce ever smaller devices, but will soon reach their fundamental limits. A promising alternative route to even smaller functional systems with nanometre dimensions is the autonomous ordering and assembly of atoms and molecules on atomically well-defined surfaces. This approach combines ease of fabrication with exquisite control over the shape, composition and mesoscale organization of the surface structures formed. Once the mechanisms controlling the self-ordering phenomena are fully understood, the self-assembly and growth processes can be steered to create a wide range of surface nanostructures from metallic, semiconducting and molecular materials.

In his classic talk of 1959, Richard P. Feynman pointed out¹ that there is “plenty of room at the bottom”. He predicted exciting new phenomena that might revolutionize science and technology and affect our everyday lives — if only we were to gain precise control over matter, down to the atomic level. The decades since then have seen the invention of the scanning tunnelling microscope that allows us to image and manipulate individual molecules and atoms^{2,3}. We also have access to nanostructured materials with extraordinary functional properties, such as semiconductor quantum dots and carbon nanotubes^{4,5}, and a growing understanding of how structural features control the function of such small systems.

These complementary developments are different aspects of nanotechnology, which aims to create and use structures, devices and systems in the size range of about 0.1–100 nm (covering the atomic, molecular and macromolecular length scales). Because of this focus on the nanometre scale, nanotechnology might meet the emerging needs of industries that have thrived on continued miniaturization and now face serious difficulties in upholding the trend, particularly in microelectronics⁶ and magnetic data storage⁷. But even if nanosystems and nanodevices with suitable performance characteristics are available, nanotechnology solutions will find practical use only if they are economically viable. We will need to develop methods for the controlled mass fabrication of functional atomic or molecular assemblies and their integration into usable macroscopic systems and devices.

The two basic approaches to creating surface patterns and devices on substrates in a controlled and repeatable manner are the ‘top-down’ and ‘bottom-up’ techniques⁸ (Fig. 1). The former may be seen as modern analogues of ancient methods such as lithography, writing or stamping, but capable of creating features down to the sub-100 nm range. The sophisticated tools allowing such precision are electron-beam writing, and advanced lithographic techniques that use extreme ultraviolet or even hard X-ray radiation⁹. Methods based on electron-beam writing achieve very high spatial resolution at reasonable capital costs, but operational capacity is limited by the serial nature of the process (although electron-projection methods may overcome this limitation). The next-generation production lines used by the semiconductor industry are likely to be based on X-ray

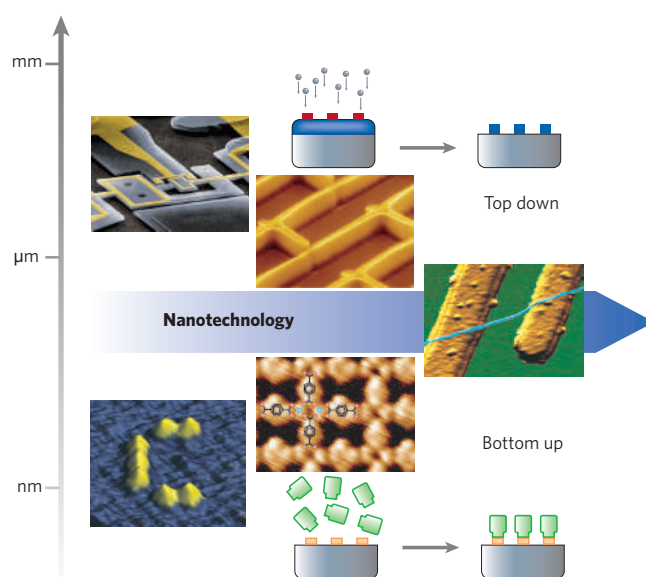


Figure 1 | Two approaches to control matter at the nanoscale. For top-down fabrication, methods such as lithography, writing or stamping are used to define the desired features. The bottom-up techniques make use of self-processes for ordering of supramolecular or solid-state architectures from the atomic to the mesoscopic scale. Shown (clockwise from top) are an electron microscopy image of a nanomechanical electrometer obtained by electron-beam lithography⁹², patterned films of carbon nanotubes obtained by microcontact printing and catalytic growth⁹³, a single carbon nanotube connecting two electrodes⁹⁴, a regular metal-organic nanoporous network integrating iron atoms and functional molecules⁷⁸, and seven carbon monoxide molecules forming the letter ‘C’ positioned with the tip of a scanning tunnelling microscope (image taken from <http://www.physics.ubc.ca/~stm/>).

lithography, which allows parallel processing. But the upgrade will require huge investments and extensive equipment development, to deal with the need for vacuum environments, short-wavelength optics, radiation sources and so on.

¹Institut de Physique des Nanostructures, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland; ²Departments of Chemistry and Physics & Astronomy, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada; ³Max-Planck-Institut für Festkörperforschung, Heisenbergstrasse 1, D-70569 Stuttgart, Germany.

Considerable efforts have also been invested in developing and exploring alternative top-down patterning methods. A particularly versatile, rapid and low-cost technique is microcontact printing¹⁰, which uses soft and hard stamps to transfer patterns with feature sizes above 100 nm onto a wide range of substrates; however, it becomes increasingly demanding for smaller feature sizes. Ultimate precision is achieved with scanning probe techniques, which are now an established (albeit cumbersome) method for the direct writing and positioning of individual atoms³. Prototype scanning force arrays that operate massively in parallel and thus multiply throughput have recently been developed¹¹, but scanning probe methods seem unlikely to find industrial use in the near future.

Top-down methods essentially ‘impose’ a structure or pattern on the substrate being processed. In contrast, bottom-up methods aim to guide the assembly of atomic and molecular constituents into organized surface structures through processes inherent in the manipulated system. Here, we outline how self-organized growth and self-assembly at well-defined surfaces (some of which may have been created using top-down methods) can serve as an efficient tool for the bottom-up fabrication of functional structures and patterns on the nanometre scale. We focus on atomic-level investigations and highlight what we regard as particularly informative illustrations of how this approach might lead to useful nanometre-scale surface structures. A brief introduction to the elementary processes governing surface self-ordering provides a foundation for the subsequent discussion of how these processes can be tuned in metallic, semiconducting and molecular systems to obtain surface structures with desired geometric order and well-defined shapes.

Basic concepts in surface structuring

Common to all bottom-up strategies for the fabrication of nanostructures at surfaces is that they are essentially based on growth phenomena. Atoms or molecules (or both) are deposited on the substrate (in vacuum, ambient atmosphere or solution) and nanometre-scale structures evolve as a result of a multitude of atomistic processes. This is inherently a non-equilibrium phenomenon and any growth scenario is governed by the competition between kinetics and thermodynamics. We thus use the term ‘self-organized growth’ to describe autonomous order phenomena mediated by mesoscale force fields or kinetic limitations in growth processes, whereas ‘self-assembly’ is reserved for the spontaneous association of a supramolecular architecture from its molecular constituents^{12–14}. The term ‘self-organization’, in contrast to self-organized growth and self-assembly, is usually used in a different context, as it relates to dissipative structure formation in systems far from thermodynamic equilibrium¹⁵ and the initial emergence of biological macromolecules¹⁶.

The primary mechanism in the growth of surface nanostructures from adsorbed species is the transport of these species on a flat terrace (see Fig. 2), involving random hopping processes at the substrate atomic lattice^{17,18}. This surface diffusion is thermally activated; that is, diffusion barriers need to be surmounted when moving from one stable (or metastable) adsorption configuration to another. As is typical for such processes, the diffusivity D — the mean square distance travelled by an adsorbate per unit time — obeys an Arrhenius law; this holds for atoms as well as rigid organic molecules¹⁹. If we now consider a growth experiment where atoms or molecules are deposited on a surface at a constant deposition rate F , then the ratio D/F determines the average distance that an adsorbed species has to travel to meet another adsorbate, either for nucleation of a new aggregate or attachment to an already formed island. The ratio of deposition to diffusion rate D/F is thus the key parameter characterizing growth kinetics. If deposition is slower than diffusion (large values of D/F), growth occurs close to equilibrium conditions; that is, the adsorbed species have enough time to explore the potential energy surface so that the system reaches a minimum energy configuration. If deposition is fast relative to diffusion (small D/F), then the pattern of growth is essentially determined by kinetics; individual processes, notably those leading to metastable structures, are increasingly important.

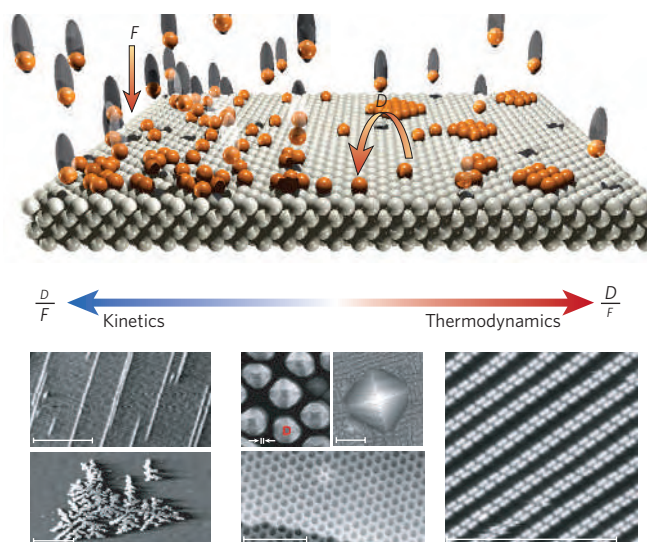


Figure 2 | Atomic-scale view of growth processes at surfaces. Atoms or molecules are deposited from the vapour phase. On adsorption they diffuse on terraces to meet other adspecies, resulting in nucleation of aggregates or attachment to already existing islands. The type of growth is largely determined by the ratio between diffusion rate D and deposition flux F . Metallic islands are controlled by growth kinetics at small D/F values. The hierarchy in the barrier of diffusing atoms can be translated into geometric order and well-defined shapes and length scales of the resulting nanostructures. The micrographs on the left-hand side show monatomic Cu chains grown on an anisotropic Pd(110) substrate (upper image) and Ag dendrites on hexagonal Pt(111) (ref. 20) (lower image). Semiconductor nanostructures are usually grown at intermediate D/F and their morphology is determined by the complex interplay between kinetics and thermodynamics. Strain effects are particularly important and can be used to achieve mesoscopic ordering. The micrographs in the centre show pyramidal and dome-shaped Ge semiconductor quantum dots grown on Si(100)⁹⁵ (upper right and upper left panels, respectively) and a boron nitride nanomesh on Rh(111)⁹⁶ (lower panel). To allow for supramolecular self-assembly based on molecular recognition, conditions close to equilibrium are required (large D/F values, or post-deposition equilibration). The micrograph on the right shows as an example a supramolecular nanograting of rod-like benzoic acid molecules on Ag(111). It consists of repulsively interacting molecular twin chains, which are stabilized by intermolecular hydrogen bonds²⁶. Scale bars, 20 nm in every image.

Low-temperature growth of metal nanostructures on metal surfaces is the prototype of kinetically controlled growth methods. Metal bonds have essentially no directionality that can be used to direct interatomic interactions. Instead, kinetic control provides an elegant way to manipulate the structure and morphology of metallic nanostructures. On homogeneous surfaces, their shape and size are largely determined by the competition between the different movements the atoms can make along the surface, such as diffusion of adatoms on surface terraces, over steps, along edges and across corners or kinks. Each of these displacement modes has a characteristic energy barrier, which will to a first approximation scale with the local coordination of the diffusing atom: the diffusion of an atom over a terrace will have a lower energy barrier than diffusion along an edge or crossing of a corner, and descending an edge is often an energetically more costly process than terrace diffusion. A given material system thus has a natural hierarchy of diffusion barriers associated with these different atomistic processes. This makes it possible to shape growing aggregates by selective activation or suppression of particular diffusion processes through external growth parameters (temperature and metal deposition flux) and through the choice of a substrate with appropriate symmetry²⁰. Judicious tuning of the relative importance of different diffusion processes has allowed on-demand fabrication of a host of metal nanostructures, ranging from small compact uniform clusters and large

faceted islands to fractals, dendrites and atomically thin chains^{17,18,20–22}.

For metals, then, formation of surface structures can be controlled only by controlling the complex kinetics of the different diffusion processes at play, given the high energies associated with covalent bond formation and the limited information for spatial organization. In contrast, molecules that can participate in weak and directional non-covalent bonds may be programmed to form desired supramolecular structures²³. The basic concepts ruling the self-assembly of three-dimensional supramolecular structures can also guide the assembly of adsorbed molecules into low-dimensional supramolecular systems that show a high degree of order on the nanometre scale. The approach does not require control over a hierarchy of activated diffusive motions, but operates near equilibrium conditions where the D/F ratio is a circumstantial parameter. Ordering may occur after deposition, and in favourable situations self-correction through the elimination of transiently formed defective structures is possible. The parameters crucial for this type of structure control are the surface mobility of molecules, their lateral interactions and their coupling to the surface atomic lattice. These depend on the chemical nature of the system and the atomic environment and symmetry of the substrate^{19,24,25}, all of which can be used to tune the delicate balance of lateral interactions and molecule–substrate coupling in order to steer supramolecular organization towards the desired structures²⁶.

If surface-supported nanostructures are to be organized on the mesoscale (10–1,000 nm), then at least some of the forces used for that purpose need to act over length scales comparable to the desired feature size; that is, they must be much longer-ranging than atomic distances. Such long-range forces can arise from various physical effects, including elastic and electrostatic interactions. Elastic forces are generally relevant to surfaces and epitaxial systems, given that atoms on a surface or in an epilayer are always under stress, even in the case of pristine surfaces or homoepitaxial systems²⁷. The resultant forces typically give rise to regular two-dimensional strain relief or reconstruction patterns^{21,22,28,29} with feature sizes of 2–20 nm, which can then serve as templates to control the growth of further patterns. In heteroepitaxial systems, the elastic energy associated with the inherent lattice mismatch of the materials can induce not only such lateral ordering, but also three-dimensional aggregation. ‘Stranski–Krastanow growth’, in which three-dimensional islands form spontaneously above a critical film thickness, is a well-established method for creating semiconductor quantum dots³⁰. This nanostructure formation process is driven by thermodynamics (that is, strain relief overcompensates for the increase in surface energy associated with the transition to three-dimensional growth), but the resulting structures are usually metastable and their exact shape, size and composition result from a delicate interplay between thermodynamic and kinetic effects.

Magnetism at the physical limit

Gordon Moore observed in 1965 that improved fabrication technologies resulted in the doubling of the number of silicon field-effect transistors per unit area roughly every 18 months. ‘Moore’s law’, which has achieved almost cult status, still describes with remarkable precision the advances in complementary metal oxide semiconductor (CMOS) technology that continue to increase information processing speeds. But the exponential growth of the information industry relies just as much on improvements in data storage, which uses small regions of ferromagnetic material with opposite magnetization to store ‘zeros’ and ‘ones’ in a hard disk. The continued downscaling of these storage domains outperforms even the stunning development of CMOS technology. During the past decade, storage density has doubled almost every 12 months and has reached 100 Gbit per square inch today. But as in the case of CMOS technologies, the drive for further miniaturization faces fundamental physical limits^{6,7}. The decrease in ferromagnetic domain size is accompanied by a decrease in the magnetic anisotropy energy K , which prevents spontaneous changes in magnetization direction. For very small domains, K is comparable to thermal

energies so that thermal fluctuations can randomly flip the magnetization direction. This renders the domains superparamagnetic, with all stable magnetic order lost. The effect can be quantified by considering that for a single domain, the time for reversal of the magnetization orientation due to thermal fluctuations follows an Arrhenius law of the type $\tau = \tau_0 \exp(nK/k_B T)$ (here τ_0 is a pre-factor of the order of 10^{-9} s, n the number of atoms in the domain, k_B the Boltzmann constant). For magnetic anisotropy energies of 40 μeV per atom (a characteristic value for bulk hexagonally close-packed cobalt) and a typical stability requirement of $\tau > 10$ years, magnetically stable nanostructures thus need to contain roughly $n \approx 10^5$ atoms. In today’s recording media, several hundred to a thousand of such individual domains are needed to realize one magnetic bit that can be reliably written and read, with high signal-to-noise ratio. Clearly, if miniaturization is to result in further increased storage capacities, we need to extend or avoid the superparamagnetic limit.

An obvious strategy is to develop new materials and structures with a substantially higher anisotropy energy K . A useful pointer is that K depends on spin–orbit interactions and on orbital magnetic moments, and hence on the precise atomic structure of magnetic materials^{31,32}. The orbital magnetic moment m_L is particularly sensitive to the local atomic configuration and influences the magnetocrystalline anisotropy. In bulk materials m_L is largely quenched through the crystal field; but in low-dimensional nanostructured materials, the reduced symmetry of the electron wavefunctions can result in strongly anisotropic orbital magnetization that will boost the magnetocrystalline contribution to the anisotropy energy. This effect should be particularly pronounced for atomic-scale structures with constituent atoms that have a reduced average coordination, which in turn results in m_L values approaching those typically seen for free atoms³³. The effect will be significant for structures that range in size from the single adatom to clusters composed of a few atoms to several tens of atoms at most, and it is difficult to envisage their efficient production without the use of bottom-up fabrication methods^{34,35}.

Judiciously used, self-organized metal growth on surfaces can produce a variety of useful nanoscale patterns with high densities in a fast, parallel process²⁰. For example, the step decoration method^{36,37} allows ready formation of uniform arrays of cobalt chains on a Pt(997) surface. The method makes use of the fact that step edges act as preferential nucleation sites for the deposited cobalt atoms, because of the increased coordination experienced at step sites relative to terrace sites. Growth in a well-defined temperature range then results in uniform cobalt chains of monatomic width over the entire sample, forming dense arrays of parallel one-dimensional nanowires. By adjusting the cobalt coverage on the surface and the average step spacing of the platinum surface, width and separation of the nanowires can be independently controlled. Similarly, bimetallic islands containing a Pt core and Co rim are readily obtained by Pt deposition and annealing to create the core islands, followed by Co deposition and annealing to create the rim³⁸.

Such surface-supported cobalt nanostructures, when examined, reveal how magnetic properties change as the size of the structures is reduced to a few atoms^{39,40}. In the case of two-dimensional cobalt clusters on Pt(111), orbital moment and magnetic anisotropy energy increase markedly as the cluster size decreases (Fig. 3a). The orbital moment increases from $m_L = 0.3 \mu_B$ (where μ_B is the Bohr magneton) for clusters composed of about 10–15 atoms, to 0.59 μ_B for tetramers and 0.78 μ_B for trimers, and to a maximum value of 1.1 μ_B for a single adatom. The latter’s orbital moment is more than seven times the bulk value. The magnetic anisotropy energy K shows a similar trend (Fig. 3a), reaching values as large as 9.3 meV for a single adatom⁴⁰. Atomic-scale Co nanostructures may thus have K values up to two orders of magnitude larger than that of bulk hexagonal close-packed (h.c.p.) cobalt. Typical magnetic recording materials such as Co/Pt multilayers and even the permanent magnet SmCo_5 also have significantly lower K values ($K \approx 0.3 \text{ meV atom}^{-1}$ and $K \approx 1.8 \text{ meV atom}^{-1}$, respectively) than the cobalt nanostructures. This pronounced size depen-

dence of magnetic properties arises from the low coordination of Co atoms in the atomic-scale nanostructures, which can reduce d -state hybridization and the crystal field potential that is produced by the electric field of neighbouring lattice atoms. We expect that detailed insight into the effects of coordination number on both m_L and K will open new avenues for the design of nanostructures with promising magnetic properties. For example, the magnetic anisotropy of two-dimensional (2D) bimetallic islands with Pt core and Co rim is entirely determined by the rim of Co atoms at the perimeter, which accounts for extreme anisotropy energies³⁸.

Two-dimensional Co nanoparticles are superparamagnetic down to the lowest temperatures. In contrast, monatomic Co chains containing on average 80 atoms (Fig. 3b, c) are ferromagnetic³⁹ at 4.2 K. The observation of a paramagnetic response at 45 K implies the

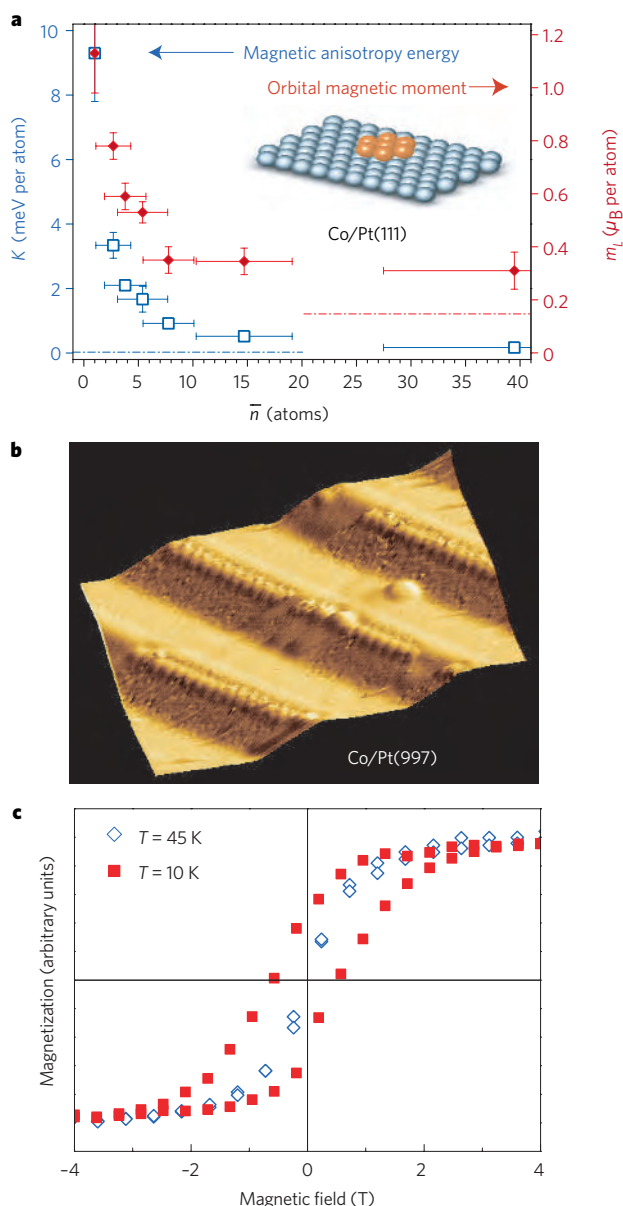


Figure 3 | Magnetism at the spatial limit. **a**, Magnetic anisotropy energy K (squares) and orbital magnetic moment m_L (diamonds) of Co atoms and two-dimensional clusters at the Pt(111) surface as a function of size n . The dashed lines represent the K and m_L values for bulk h.c.p. Co (blue and red, respectively)⁴⁰. **b**, Scanning tunnelling microscopy image of monatomic cobalt chains decorating the steps of a regularly stepped platinum surface. The average distance between neighbouring chains is 2 nm. **c**, On cooling below 15 K, Co blocks become ferromagnetic, indicated by the opening up of a hysteresis loop in the magnetization curve³⁹.

absence of single-domain 1D ferromagnetic coupling, but the shape of the magnetization curve reveals the onset of short-range magnetic order, with spins coupling into local blocks of roughly 15 atoms. Long-range order is forbidden in infinite 1D systems at true equilibrium, yet it may exist in finite systems over relatively short timescales. In the case of the monatomic Co chain, a transition into a long-range ferromagnetically ordered state occurs at 15 K and is evident from the hysteresis in the magnetization curve (see Fig. 3c). This behaviour is due to the low coordination of the Co atoms on the vicinal Pt substrate, which results in a large anisotropy energy of 2.1 meV atom⁻¹ that locks the magnetization of each spin block along the easy axis of the system. On the timescale of the experiment, ferromagnetic coupling thus effectively extends over the entire chain array⁴¹ and gives rise to the smallest elemental magnet yet fabricated.

The magnetic behaviour seen in Co nanostructures is not only relevant for our fundamental understanding of low-dimensional magnetism, but has important implications for magnetic data storage technology. That is, the increase in magnetic anisotropy energies by more than two orders of magnitude, relative to the values seen in more traditional transition-metal systems, implies that a few hundred cobalt atoms might suffice to realize a stable magnetic bit at room temperature.

Semiconductor artificial atoms

In metallic nanostructures, because of the effects of coordination, every atom 'counts' with respect to the magnetic properties. For semiconductor materials, functional properties tend to be less sensitive to the exact number of constituent atoms: desired quantum effects already arise for structures with dimensions of 10–100 nm and containing somewhere between 10^3 and 10^6 atoms in a crystalline lattice. In this size range, the energy spectrum of electrons and holes confined in all three dimensions within these 'quantum dots' becomes discrete and in many ways similar to the spectrum of atoms⁴. Still, quantum dots are very much solid-state nanostructures, and their energy spectrum, which controls many of the physical properties of interest, can be adjusted over a wide range by tuning composition, size, lattice strain and morphology. These features make semiconductor quantum dots attractive for the design and fabrication of new electronic, magnetic and photonic devices and other functional materials.

Semiconductor quantum dots are often prepared as colloidal nanocrystals (see the review in this issue by Yin & Alivisatos, p. 664) but here we will focus only on semiconductor quantum dots supported on surfaces or embedded in solids. These systems can be prepared by using a wide range of methods, including lithography, etching and site-selective implantation⁴². But fabrication methods based on self-organized growth at surfaces are particularly attractive because they yield quantum dots with virtually no interface defects that might adversely affect performance, and because they can produce particularly small structures with widely separated energy levels that are essential for room-temperature operation. The approach can also produce high-density quantum dot structures in a fast, parallel process that is compatible with existing semiconductor technology and therefore permits mass fabrication and high levels of integration. However, to use self-organized growth effectively for quantum dot fabrication, we need detailed insight into the nucleation and growth processes involved, so as to tune the dot properties precisely and control their intrinsic statistical inhomogeneity.

As so many physical properties depend on quantum dot size and shape, it is essential to know the actual morphology of the 3D semiconductor islands that form on deposition of atoms from the gas phase, and to know how they evolve during post-growth treatments. But even though these systems have been intensely studied for more than a decade and many of their electronic and optical properties characterized, the structure of nucleated semiconductor islands and their subsequent morphological evolution remain incompletely understood. It is therefore encouraging that a common framework⁴³ can describe quantum dots that develop in the two most studied model

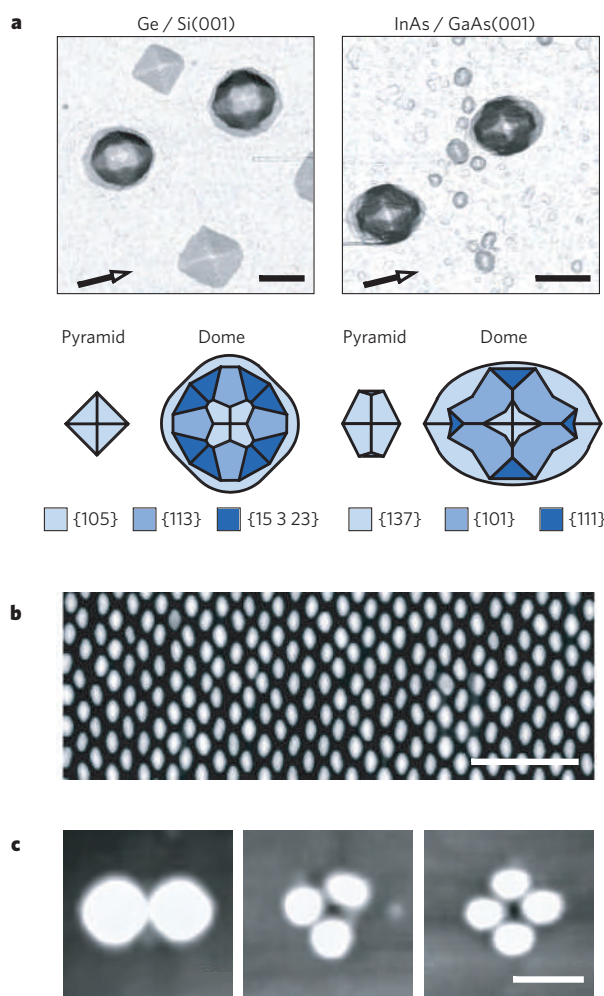


Figure 4 | Semiconductor quantum structures. **a**, Scanning tunnelling microscope images of pyramid and dome islands for the two main representative systems in semiconductor lattice-mismatched heteroepitaxy. The corresponding schematic structural models are also shown⁴³. **b**, Atomic force topography of a regular array of InGaAs quantum dots reflecting self-organized growth on a prestructured GaAs(001) substrate⁵⁰. **c**, Lateral quantum dot molecules grown in the InAs/GaAs(001) system. Bi-, tri- and quad-molecules can be produced by adjusting the substrate temperature and amount of deposited material⁶⁰. Scale bars correspond to 50 nm in **a** and **c**, and to 500 nm in **b**.

systems: during the growth of Ge on Si(001) and during the growth of InAs on GaAs(001). The two types of quantum dots are both produced in the Stranski–Krastanow growth mode, with defect-free but strained 3D islands forming spontaneously on top of a thin wetting layer during lattice-mismatched heteroepitaxial growth. In both systems, only two discrete, well-defined families of islands develop: small islands that are bounded by one type of shallow facets and referred to as pyramids, and larger, multi-faceted islands that are characterized by steeper facets and referred to as domes (see Fig. 4a). When overgrowing the initially formed islands with the substrate material (Si and GaAs, respectively) to create the actual quantum dots, the capping process in both systems involves extension of the shallow facets at the expense of the steeper ones and a considerable reduction in island height. These experimental observations confirm theoretical predictions⁴⁴ that common, well-defined island shapes occur during growth and evolution, independent of the specific material system considered. It might therefore be possible to develop a common framework to explain at least qualitatively island growth and evolution for many material combinations that follow the Stranski–Krastanow growth mode. We expect the availability of such a universally applicable descriptive model to have

considerable impact on our ability to design and engineer quantum dot structures.

In addition to controlling the properties of individual quantum dots, many of the applications so far envisaged also require precise arrangement of these structures into ordered arrays. For example, regular arrangement is obvious for systems or devices that require the addressing or coupling of individual quantum dots or the further processing of quantum dot signals, as in the case of single-electron⁴⁵, single-photon⁴⁶ and quantum computation⁴⁷ devices. Similarly, uniformity in position and spacing is critical for applications that make use of quantum dot ensembles, and where overall device performance depends on the mutual interaction between the individual dots (as in the case of cellular automata based on quantum dots⁴⁸, discussed below). But a high degree of lateral order has the additional advantage that formation of these structures usually ensures high uniformity in quantum dot properties, as statistical fluctuations are greatly suppressed if each dot experiences the same local environment during growth.

Quantum dots may be laterally organized using self-ordering processes that are mediated by elastic interactions, or using patterned substrates to direct their growth. The former approach depends on the same forces that induce the spontaneous formation of islands in the Stranski–Krastanow growth mode, and will drive the ordering of islands on length scales of the same order of magnitude as the size of the islands⁴⁹. But it is almost impossible to obtain defect-free mesoscopic quantum dot arrangements based on this approach only. In contrast, highly regular structures are readily obtained by using electron-beam or optical lithography (top-down methods) first to create patterned substrates, which then serve as templates to direct the self-organized Stranski–Krastanow growth of three-dimensional semiconductor islands (a bottom-up method). The artificial periodic modulations of the surface are thus translated into perfect quantum dot arrays^{50,51}, as illustrated by the example shown in Fig. 4b. A further advantage of this approach is that it allows the resulting quantum dot structures to be connected to larger structures for integration into complex devices.

Efforts to controllably fabricate and characterize semiconductor quantum dots have mainly been driven by the desire to develop systems that take advantage of their extremely small dimensions and low power dissipation. For example, the performance of lasers can be substantially improved by using quantum dots as the active medium⁵². The tunable and discrete energy levels typical of quantum dots mean that the choice of emitted wavelength can be adjusted with unprecedented flexibility; the small active volume permits laser operation at low power, high frequency and low threshold currents that are independent of the working temperature. Information technology is another field in which the properties of quantum dots might prove attractive. One example is a cellular automaton⁴⁸ in which binary information is encoded in the configuration of charge distributed among the quantum dots and interaction between the dots provided by Coulomb interactions. Such an ensemble, with appropriately designed dot arrangement, is an essentially classical device that can reproduce the effect of wires and logic gates and implement any complex Boolean operation⁵³. But quantum dots are also attracting interest for quantum computation applications, where information is encoded using quantum two-state systems ('qubits') that can be prepared in a superposition of the two states and thus enable dramatic increases in computing capabilities^{54–56}. The basic building block is a two-qubit quantum gate that performs unitary operations on one qubit depending on the (quantum) state of the other⁵⁷. Such a gate can be realized using quantum dot molecules; that is, sets of quantum-mechanically coupled quantum dots whose charge carrier wavefunctions are delocalized over the entire structure. The first semiconductor quantum dot molecules with controlled separation between the dots have been fabricated using top-down methods⁵⁸, although their sensitivity to thermal perturbations precludes any scope for use in 'real world' devices. Decreasing the size of the quantum dots should allow

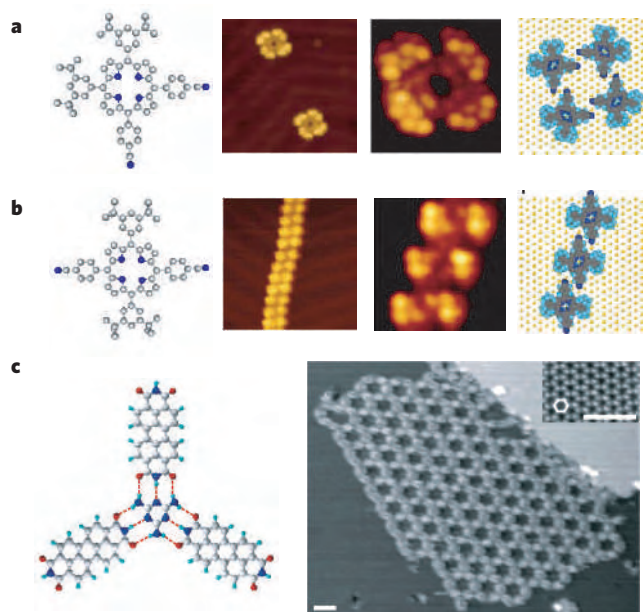


Figure 5 | Steering self-assembly of supramolecular nanostructures using hydrogen-bonding. **a, b**, Porphyrins substituted with two functional cyanophenyl moieties in a *cis* or *trans* configuration⁶⁷. **a**, The *cis* species assembles in compact clusters of four molecules. **b**, With the *trans* species linear molecular chains are obtained. Mesoscopic ordering is in both cases dictated by the preferential attachment at the elbows of the chevron reconstruction of the Au(111) substrate used, visible as weak corrugation lines. Imaged area at left (right) is 20 nm² (5.3 nm²). **c**, The complementary tectons perylene tetracarboxylic di-imide and melamine form a trigonal motif, where each intermolecular linkage is stabilized by three hydrogen bonds. This repeat unit gives rise to the regular nanoporous honeycomb layers fabricated on the hexagonal Ag-passivated Si(111) substrate shown in the inset⁷⁵. Scale bars, 3 nm.

room-temperature operation. But controlled formation and precise arrangement of such small structures are challenging and likely to require self-organized growth, as illustrated by a strategy that makes use of spontaneous alignment of stacked quantum dot layers to form vertical quantum dot molecules⁵⁹. The selective addressing of specific quantum gate parts is likely to be more feasible using laterally coupled quantum dot molecules, which have now also been produced through self-organized growth^{60,61}. Figure 4c shows quantum dot molecules containing two, three and four dots; these were created in InAs/GaAs(001) using an elaborate growth–overgrowth–etching–regrowth procedure that is primarily based on self-organized growth⁶⁰. The electronic properties of these quantum dot molecules remain to be explored, but their structural characteristics and the ability to combine them into highly ordered arrays are promising.

Supramolecular engineering

The range of functional nanometre-sized structures that can be fashioned from metallic or semiconducting materials through self-organized growth is inevitably somewhat limited, in that design and fabrication methods need to be based on the functional and structural features inherent in these materials. This contrasts with the construction of molecular nanoscale structures and patterns: the power of chemical synthesis provides access to a potentially vast range of functionally and structurally diverse building blocks ('tectons'), which can be linked through different types of relatively weak, non-covalent interactions (predominantly hydrogen bonds and metal–ligand interactions) to yield organized supramolecular architectures with tailor-made properties^{14,23,62}. But although much is known about how supramolecular chemistry — the chemistry of the intermolecular non-covalent bond — can be tuned to create desired supramolecular crystals or supramolecular compounds in solution, this knowledge cannot be directly translated to guide the assembly

of adsorbed molecules into larger surface structures^{63,64}. To do so, the influence of the substrate atomic lattice and substrate electronic structure on non-covalent bonds needs to be fully understood. For example, the substrate used will often alter the electronic properties of adsorbed ligands so that solution-based coordination chemistry concepts cannot be applied without appropriate modification⁶⁵. Interactions between adsorbed molecules and their substrate may also perturb the surface-state free electron gas in metallic substrates, which in turn can influence how adsorbates are arranged on the surface⁶⁶. More direct effects are that substrates may be reactive and chemically modify the functional moieties of the adsorbed building blocks, and the fact that topological surface features can influence interactions between adsorbed species for geometric reasons. These various effects make it obviously difficult to translate supramolecular concepts developed for crystals or solutions; but they can be used as tools for steering non-covalent interactions through the choice of templates with appropriate symmetry, surface patterns or chemical functions.

Planar molecules with extended π -systems have found particularly wide use because they tend to bond to surfaces in a flat-lying geometry, which allows functional groups at the molecular periphery to approach each other easily and to engage in non-covalent interactions. Provided the molecules retain mobility on the substrate with their functional groups not obstructed by the surface, supramolecular surface structures readily form as a result of two-dimensional self-assembly: the lateral coupling of suitably designed tectons. The tunability of supramolecular surface patterns through tecton design is illustrated in Fig. 5a, b⁶⁷: the exact position of two cyanophenyl substituents at the periphery of a porphyrin core steers the intermolecular hydrogen bonding that provides lateral coupling and hence dictates the structure of the assemblies formed on a gold surface. Whereas the *cis* configuration gives rise to discrete clusters made up of four molecules (Fig. 5a), the *trans* configuration produces extended one-dimensional supramolecular chains (Fig. 5b). The spatial distribution of these clusters and chains is dictated by the surface pattern of the gold surface, which in the case of the Au(111) surface used results from a chevron reconstruction²⁸. A similar influence on supramolecular ordering is seen in the case of 1-nitronaphthalene⁶⁸, emphasizing that patterned substrates are generally useful to guide the formation of low-dimensional molecular nanostructures^{69,70}.

A systematic study of the self-assembly behaviour of 4-[*trans*-2-(pyrid-4-yl-vinyl)]benzoic acid (PVBA) illustrates how the materials characteristics and symmetry of the substrate can affect the subtle balance between intermolecular interactions and molecule–surface interactions. The rod-like PVBA molecule, which contains a benzoic acid head group and a pyridyl tail group, self-assembles through head-to-tail hydrogen bonding^{26,71}. If metallic palladium is used as substrate, molecule–substrate coupling is strong and dominates over intermolecular interactions; this prevents the formation of regular surface patterns. On close-packed noble-metal surfaces, the PVBA molecules are more mobile and able to assemble into highly regular, one-dimensional supramolecular arrangements resembling 'nanogratings' (see for example the scanning tunnelling microscope image reproduced in Fig. 2)²⁶. Each stripe in the nanograting consists of two discrete chains of hydrogen-bonded PVBA molecules. The chains making up one stripe are held together through weak inter-chain hydrogen bonds, and the patterning of the stripes relative to each other appears to be due to interchain repulsions. The stripes are each about 1 nm wide, but the periodicity of the grating can be tuned from about 2 to 10 nm by controlling how much PVBA is deposited, or by taking advantage of the fact that they preferentially assemble at dislocation arrays of reconstructed substrates such as Au(111)^{71,72}. Surface feature control at the molecular length scale is beyond the limits of current lithographic techniques. But for the self-assembled patterns to find practical use, methods need to be developed to transform the molecular arrangements into more rigid structures while retaining their precise spatial organization.

Extended 2D open network structures have been created using a number of different systems. One example is trimesic acid (TMA, $C_6H_3(COOH)_3$), in which regular dimerization of the carboxyl groups present results in an open network structure that reflects the molecule's three-fold symmetry^{73,74}. This pattern is also encountered in organic TMA crystals, illustrating that in favourable cases motifs from organic solids can be replicated on surfaces. Another extended network structure is based on the classic H-bond motif of the melamine–cyanuric acid system, with perylene tetracarboxylic diimide serving as linear linker and melamine as trigonal connector. As illustrated in Fig. 5c, self-assembly through triple H-bond coupling yields a 2D bimolecular honeycomb network⁷⁵. The high degree of order obtained on a substrate with appreciable surface corrugation (Ag-passivated Si(111)) emphasizes that even though individual intermolecular couplings may be weak, multiple weak linkages nevertheless enable stable and regular assemblies to form. Network stability is an essential feature if nanoporous surface patterns are used as templates to guide the formation of subsequent layers, as demonstrated by the formation of regular C_{60} arrays on the 2D bimolecular honeycomb arrangement⁷⁵.

Metal–ligand interactions are generally stronger than hydrogen bonds and thus result in more robust entities. Moreover, the incorporation of metal centres increases the scope of the functional properties of the nanoarchitectures and allows us to use design strategies based on metal-directed assembly^{76,77}. This makes the controlled fabrication of surface structures based on metal–organic coordination appealing, but the approach can be challenging. Difficulties arise from the tendency of deposited metal atoms to interact strongly with the substrates used, in extreme cases resulting in surface reconstruction or alloying. Consequently the deposition sequence and substrate temperature have to be carefully controlled to avoid spurious effects and to achieve the formation of regular nanosystems.

We illustrate the principles underlying the interaction of organic linker molecules and transition metals at surfaces with the coordination behaviour of benzenepolycarboxylic acid species and iron (Fe) atoms on a copper Cu(100) surface. This system forms a variety of two-dimensional surface-supported open networks. The basic architectural motif of these networks depends on the relative concentrations of metal atoms and ligand molecules, with careful tuning of this ratio resulting in mononuclear metal-carboxylate clusters, 1D coordination polymers or fully connected 2D networks. The mononuclear complexes (Fig. 6a) are obtained⁷⁸ on depositing about 0.3 Fe atoms per linear terephthalate linker molecule (TPA, $C_6H_4(COO)_2$). As indicated in the scheme, a central Fe atom is coordinated to four TPA molecules through Fe-carboxylate bonds, with the resultant $Fe(TPA)_4$ complexes organized in a (6×6) unit cell with respect to the substrate. The complexes form a highly ordered array that covers entire terraces of the substrate, with this perfect long-range organization suggested^{78,79} to result from weak hydrogen-bonding interactions between the complexes. The individual Fe centres are thus arranged in a perfect square lattice with a 15 Å periodicity. Clearly, any attempt to position large-scale arrays of single Fe atoms in a similar way using a top-down technique would be prohibitively time-consuming, if not impossible.

Increasing the amount of deposited metal to about 1 Fe atom per linker molecules results in networks of polymeric ladder structures, as illustrated in Fig. 6b for the system based on trimellate as linker⁸⁰. This network frequently covers entire substrate terraces and constitutes a regular array of open nanocavities, each with an effective opening of $(3 \times 10) \text{ Å}^2$. Increasing the metal concentration further, to about two Fe atoms per linker molecule, yields fully interconnected 2D metal–organic networks with complete two-dimensional reticulation^{78,80}. Examples are shown in one of the insets of Fig. 1 (using TPA as linker⁷⁸) and in Fig. 6c (using as linker a longer analogue of TPA, 4,1',4'',1''-terphenyl-1,4''-dicarboxylic acid or TDA⁸¹). Both networks are thermally robust. Because the networks 'compartmentalize' the copper substrate into nanometre-sized cavities, they can steer the

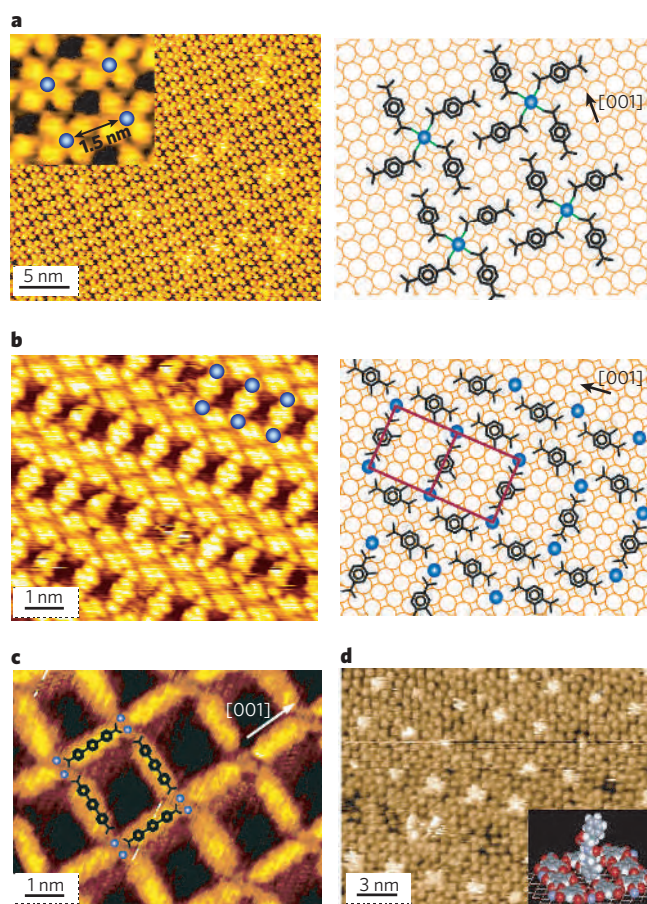


Figure 6 | Metallosupramolecular assembly of low-dimensional Fe-carboxylate coordination systems on a square Cu(100) substrate. **a**, The mononuclear $Fe(TPA)_4$ complexes are stabilized by metal–ligand interactions. Their perfect ordering in a (6×6) square array with periodicity 15 Å is mediated by substrate templating and weak intercomplex hydrogen bridges⁷⁸. **b**, One-dimensional Fe-trimellate coordination polymer with a higher coverage ratio of Fe per tecton. In the ladder structure indicated, with a (4×4) repeat unit, there is a continuous 1D Fe-carboxylate linkage framing open cavities⁸⁰. **c**, Using a linear terphenyl dicarboxylate species as linker with increased length, regular 2D reticulated coordination networks with nanometre-sized cavities are obtained⁸¹. **d**, Site-selective uptake of L, L-diphenylalanine peptide molecules in a 2D nanoporous Fe-carboxylate architecture⁸⁴; a model for the dipeptide guests in an upright position is depicted in the inset (S. Stepanow, N. Lin, J.V.B. and K.K., unpublished observations).

organization of subsequently deposited molecules, as has been illustrated with C_{60} molecules⁸¹.

As might be expected, substrate and linker symmetry play an important role in 2D supramolecular engineering of metal–organic surface structures. By replacing the Cu(100) surface with its square symmetry by the anisotropic Cu(110) surface, the 1D anisotropy of the substrate can effectively be transferred to the resulting coordination compounds. Trimesic acid with Cu and Fe on the (110) surface is found to form strictly linear metal–organic coordination chains⁸². The local linker geometry is of equal importance. For instance, whereas TMA linkers with three-fold molecular symmetry can form mononuclear $Fe(TMA)_4$ complexes that resemble their terephthalate $Fe(TPA)_4$ counterparts⁸³, the complexes cannot assemble into perfectly square arrays but form instead an extended coordination network with a regular arrangement of nanocavities⁸⁴. These nanocavities provide well-defined reaction spaces and can even be used as selective receptors for biomolecules. This is as demonstrated in Fig. 6d where the L, L-diphenylalanine peptide species is used as molecular guest in a two-dimensional metal-carboxylate host-system (S. Stepanow, N. Lin, J.V.B. and K.K.,

unpublished observation). The design principles established for 3D carboxylate framework reticular synthesis^{85,86} can thus also guide the design of 2D supramolecular metal–organic systems. That is, careful selection of suitable linker structures (in conjunction with appropriate metals) provides an effective strategy for adjusting the size of the cavities or pores present in the 2D assemblies, and the chemical functionalities lining the ‘walls’ of the pores are determined by the characteristics of the side groups present in the linker. Given the immense wealth and scope of reticular chemistry, a wide range of surface-supported structures with pores of different sizes and chemical characteristics should in principle be accessible. These might find use in the fabrication of patterned surface templates, for the control of host–guest chemistry involving surface structures, or in heterogeneous catalysis in which reactants can interact with the substrate and the supramolecular metal–organic surface structures.

Outlook

Self-organized growth and self-assembly at surfaces can serve as an efficient and versatile tool for creating low-dimensional nanostructures. It offers exquisite control over feature size and organization on the atomic and mesoscopic length scales. We believe that these process characteristics, in combination with the ability to produce high-density structures in a fast and parallel fashion, are essential requirements for any nanofabrication methodology that aims to contribute to the quest for further miniaturization in the microelectronics industry and elsewhere. However, even though processes that make use of self-ordering growth have already yielded systems with intriguing functional properties, many challenges still need to be addressed before such strategies find wide practical use. For example, the incorporation of nanostructures into more complex organized architectures and their effective interfacing to the macroscopic world are vital for any applications. We would expect that this can be achieved by combining bottom-up and top-down techniques, with the former providing ready access to features with sizes below 10 nm, and the latter allowing for integration of these structures into larger functional systems. This general approach should also result in new materials and devices that might find use beyond the applications traditionally targeted by miniaturization efforts, particularly when it is guided by new insights into the physics of small systems or combined with chemical^{35,41,87,88} and biological^{89–91} bottom-up methods. ■

1. Feynman, R. P. There's plenty of room at the bottom. *Eng. Sci.* **23**, 22–36 (1960).
2. Binnig, G. & Rohrer, H. Scanning tunneling microscopy — from birth to adolescence. *Rev. Mod. Phys.* **59**, 615–625 (1987).
3. Eigler, D. M. & Schweizer, E. K. Positioning single atoms with a scanning tunnelling microscope. *Nature* **344**, 524–526 (1990).
4. Kastner, M. A. Artificial atoms. *Phys. Today* **46**, 24–31 (1993).
5. Dekker, C. Carbon nanotubes as molecular quantum wires. *Phys. Today* **52**, 22–28 (1999).
6. Bohr, M. T. Nanotechnology goals and challenges for electronics applications. *IEEE Trans. Nanotechnol.* **1**, 56–62 (2002).
7. Thomson, D. A. & Best, J. S. The future of magnetic data storage technology. *IBM J. Res. Dev.* **3**, 311–321 (2000).
8. Gates, B. D. et al. New approaches to nanofabrication: molding, printing, and other techniques. *Chem. Rev.* **105**, 1171–1196 (2005).
9. Ito, T. & Okazaki, S. Pushing the limits of lithography. *Nature* **406**, 1027–1031 (2000).
10. Xia, Y. N., Rogers, J. A., Paul, K. E. & Whitesides, G. M. Unconventional methods for fabrication and patterning nanostructures. *Chem. Rev.* **99**, 1823–1848 (1999).
11. Vettiger, P. et al. The ‘millipede’—nanotechnology entering data storage. *IEEE Trans. Nanotechnol.* **1**, 39–55 (2002).
12. Lindsey, J. S. Self-assembly in synthetic routes to devices. Biological principles and chemical perspectives: a review. *New J. Chem.* **15**, 153–180 (1991).
13. Whitesides, G. M., Mathias, J. P. & Seto, C. T. Molecular self-assembly and nanochemistry — a chemical strategy for the synthesis of nanostructures. *Science* **254**, 1312–1319 (1991).
14. Philp, D. & Stoddart, J. F. Self-assembly in natural and unnatural systems. *Angew. Chem. Int. Edn Engl.* **35**, 1154–1196 (1996).
15. Nicolis, G. & Prigogine, I. *Self-Organization in Non-Equilibrium Systems: From Dissipative Structure Formation to Order through Fluctuations* (Wiley, New York, 1977).
16. Eigen, M. Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **33**, 465–523 (1971).
17. Zhang, Z. Y. & Lagally, M. G. Atomistic processes in the early stages of thin-film growth. *Science* **276**, 377–383 (1997).
18. Brune, H. Microscopic view of epitaxial growth: nucleation and aggregation. *Surf. Sci. Rep.* **31**, 121–229 (1998).
19. Barth, J. V. Transport of adsorbates at metal surfaces: From thermal migration to hot precursors. *Surf. Sci. Rep.* **40**, 75–150 (2000).
20. Röder, H., Hahn, E., Brune, H., Bucher, J. P. & Kern, K. Building one-dimensional and two-dimensional nanostructures by diffusion-controlled aggregation at surfaces. *Nature* **366**, 141–143 (1993).
21. Brune, H., Giovannini, M., Bromann, K. & Kern, K. Self-organized growth of nanostructure arrays on strain-relief patterns. *Nature* **394**, 451–453 (1998).
22. Li, J. L. et al. Spontaneous assembly of perfectly ordered identical-size nanocluster arrays. *Phys. Rev. Lett.* **88**, 066101 (2002).
23. Lehn, J.-M. *Supramolecular Chemistry, Concepts and Perspectives* (VCH, Weinheim, 1995).
24. Gimzewski, J. K. & Joachim, C. Nanoscale science of single molecules using local probes. *Science* **283**, 1683–1688 (1999).
25. Rosei, F. et al. Properties of large organic molecules at surfaces. *Prog. Surf. Sci.* **71**, 95–146 (2003).
26. Barth, J. V. et al. Building supramolecular nanostructures at surfaces by hydrogen bonding. *Angew. Chem. Int. Edn Engl.* **39**, 1230–1234 (2000).
27. Ibach, H. The role of surface stress in reconstruction, epitaxial growth and stabilization of mesoscopic structures. *Surf. Sci. Rep.* **29**, 193–263 (1997).
28. Barth, J. V., Brune, H., Ertl, G. & Behm, R. J. Scanning tunneling microscopy observations on the reconstructed Au(111) surface — atomic structure, long-range superstructure, rotational domains, and surface defects. *Phys. Rev. B* **42**, 9307–9318 (1990).
29. Kern, K. et al. Long-range spatial self-organization in the adsorbate-induced restructuring of surfaces — Cu(110)-(2×1)O. *Phys. Rev. Lett.* **67**, 855–858 (1991).
30. Teichert, C. Self-organization of nanostructures in semiconductor heteroepitaxy. *Phys. Rep.* **365**, 335–432 (2002).
31. Bruno, P. Tight-binding approach to the orbital magnetic moment and magnetocrystalline anisotropy of transition-metal monolayers. *Phys. Rev. B* **39**, 865–868 (1989).
32. van der Laan, G. Microscopic origin of magnetocrystalline anisotropy in transition metal thin films. *J. Phys. Cond. Mat.* **10**, 3239–3253 (1998).
33. Wildberger, K., Stepanyuk, V. S., Lang, P., Zeller, R. & Dederichs, P. H. Magnetic nanostructures — 4d clusters on Ag(001). *Phys. Rev. Lett.* **75**, 509–512 (1995).
34. Himpel, F. J., Ortega, J. E., Mankey, G. J. & Willis, R. F. Magnetic nanostructures. *Adv. Phys.* **47**, 511–597 (1998).
35. Sun, S., Murray, C. B., Weller, D., Folks, L. & Moser, A. Monodisperse FePt nanoparticles and ferromagnetic FePt nanocrystal superlattices. *Science* **287**, 1989–1992 (2000).
36. Gambardella, P., Blanc, M., Brune, H., Kuhnke, K. & Kern, K. One-dimensional metal chains on Pt vicinal surfaces. *Phys. Rev. B* **61**, 2254–2262 (2000).
37. Kuhnke, K. & Kern, K. Vicinal metal surfaces as nanotemplates for the growth of low-dimensional structures. *J. Phys. Cond. Mat.* **15**, S3311–S3335 (2003).
38. Rusponi, S. et al. The remarkable difference between surface and step atoms in the magnetic anisotropy of two-dimensional nanostructures. *Nature Mater.* **2**, 546–551 (2003).
39. Gambardella, P. et al. Ferromagnetism in one-dimensional monatomic metal chains. *Nature* **416**, 301–304 (2002).
40. Gambardella, P. et al. Giant magnetic anisotropy of single cobalt atoms and nanoparticles. *Science* **300**, 1130–1133 (2003).
41. Gambardella, P. et al. Oscillatory magnetic anisotropy in one-dimensional atomic wires. *Phys. Rev. Lett.* **93**, 077203 (2004).
42. Bimberg, D., Grundmann, M. & Ledentsov, N. N. *Quantum Dot Heterostructures* (Wiley, Chichester, 1999).
43. Costantini, G. et al. Universal island shapes of self-organized semiconductor quantum dots. *Appl. Phys. Lett.* **85**, 5673–5675 (2004).
44. Daruka, I., Tersoff, J. & Barabasi, A. L. Shape transition in growth of strained islands. *Phys. Rev. Lett.* **82**, 2753–2756 (1999).
45. Warburton, R. J. et al. Optical emission from a charge-tunable quantum ring. *Nature* **405**, 926–929 (2000).
46. Michler, P. et al. A quantum dot single-photon turnstile device. *Science* **290**, 2282–2284 (2000).
47. Burkard, G., Loss, D. & DiVincenzo, D. P. Coupled quantum dots as quantum gates. *Phys. Rev. B* **59**, 2070–2078 (1999).
48. Cole, T. & Luthy, J. C. Quantum-dot cellular automata. *Prog. Quantum Electron.* **25**, 165–189 (2001).
49. Shchukin, V. A. & Bimberg, D. Spontaneous ordering of nanostructures on crystal surfaces. *Rev. Mod. Phys.* **71**, 1125–1171 (1999).
50. Heidemeyer, H., Denker, U., Müller, C. & Schmidt, O. G. Morphology response to strain field interferences in stacks of highly ordered quantum dot arrays. *Phys. Rev. Lett.* **91**, 196103 (2003).
51. Lee, H., Johnson, J. A., He, M. Y., Speck, J. S. & Petroff, P. M. Strain-engineered self-assembled semiconductor quantum dot lattices. *Appl. Phys. Lett.* **78**, 105–107 (2001).
52. Arakawa, Y. & Sakaki, H. Multidimensional quantum well laser and temperature-dependence of its threshold current. *Appl. Phys. Lett.* **40**, 939–941 (1982).
53. Nakajima, F., Miyoshi, Y., Motohisa, J. & Fukui, T. Single-electron AND/NAND logic circuits based on a self-organized dot network. *Appl. Phys. Lett.* **83**, 2680–2682 (2003).
54. Feynman, R. P. Simulating physics with computers. *Int. J. Theor. Phys.* **21**, 467–488 (1982).
55. Shor, P. in *Proc 35th Annu. Symp. Foundations of Computer Science* (IEEE, Los Alamitos, 1994).
56. Deutsch, D. Quantum theory, the Church–Turing principle and the universal quantum computer. *Proc. R. Soc. Lond. A* **400**, 97–117 (1985).
57. Barenco, A. et al. Elementary gates for quantum computation. *Phys. Rev. A* **52**, 3457–3467 (1995).
58. Schedelbeck, G., Wegscheider, W., Bichler, M. & Abstreiter, G. Coupled quantum dots fabricated by cleaved edge overgrowth: From artificial atoms to molecules. *Science* **278**, 1792–1795 (1997).
59. Bayer, M. et al. Coupling and entangling of quantum states in quantum dot molecules. *Science* **291**, 451–453 (2001).
60. Songmuang, R., Kiravittaya, S. & Schmidt, O. G. Formation of lateral quantum dot molecules around self-assembled nanoholes. *Appl. Phys. Lett.* **82**, 2892–2894 (2003).
61. Deng, X. & Krishnamurthy, M. Self-assembly of quantum-dot molecules: heterogeneous nucleation of SiGe islands on Si(100). *Phys. Rev. Lett.* **81**, 1473–1476 (1998).
62. Prins, L. J., Reinhoudt, D. N. & Timmerman, P. Non-covalent synthesis using hydrogen bonding. *Angew. Chem. Int. Edn Engl.* **40**, 2382 (2001).
63. Barth, J. V., Weckesser, J., Lin, N., Dmitriev, S. & Kern, K. Supramolecular architectures and nanostructures at surfaces. *Appl. Phys. A* **76**, 645 (2003).

64. Feyter, S. D. & Schryver, F. C. D. Two-dimensional supramolecular self-assembly probed by scanning tunneling microscopy. *Chem. Soc. Rev.* **32**, 139–150 (2003).
65. Lin, N., Dmitriev, A., Weckesser, J., Barth, J. V. & Kern, K. Real-time single-molecule imaging of the formation and dynamics of coordination compounds. *Angew. Chem. Int. Edn Engl.* **41**, 4779 (2002).
66. Lukas, S., Witte, G. & Wöll, C. Novel mechanism for molecular self-assembly on metal substrates: unidirectional rows of pentacene on Cu(110) produced by substrate-mediated repulsion. *Phys. Rev. Lett.* **88**, 028301 (2002).
67. Yokoyama, T., Yokoyama, S., Kamikado, T., Okuno, Y. & Mashiko, S. Selective assembly on a surface of supramolecular aggregates of controlled size and shape. *Nature* **413**, 619–621 (2001).
68. Böhringer, M. *et al.* Two-dimensional self-assembly of supramolecular clusters and chains. *Phys. Rev. Lett.* **83**, 324–327 (1999).
69. Otero, R. *et al.* One-dimensional assembly and selective orientation of Lander molecules on an O-Cu template. *Angew. Chem. Int. Edn Engl.* **43**, 2092–2095 (2004).
70. Clair, S., Pons, S., Brune, H., Kern, K. & Barth, J. V. Mesoscopic metallo-supramolecular texturing through hierarchic assembly. *Angew. Chem. Int. Edn Engl.* (in the press).
71. Barth, J. V. *et al.* Stereochemical effects in supramolecular self-assembly at surfaces: 1-D vs. 2-D enantiomorphic ordering for PVBA and PEBA on Ag(111). *J. Am. Chem. Soc.* **124**, 7991–8000 (2002).
72. Weckesser, J., Vita, A. D., Barth, J. V., Cai, C. & Kern, K. Mesoscopic correlation of supramolecular chirality in one-dimensional hydrogen-bonded assemblies. *Phys. Rev. Lett.* **87**, 096101 (2001).
73. Dmitriev, A., Lin, N., Weckesser, J., Barth, J. V. & Kern, K. Supramolecular assemblies of trimesic acid on a Cu(100) surface. *J. Phys. Chem. B* **106**, 6907–6912 (2002).
74. Griessl, S., Lackinger, M., Edelwirth, M., Hietschold, M. & Heckl, W. M. Self-assembled two-dimensional molecular host-guest architectures from trimesic acid. *Single Molecules* **3**, 25–31 (2002).
75. Theobald, J. A., Oxtoby, N. S., Phillips, M. A., Champness, N. R. & Beton, P. H. Controlling molecular deposition and layer structure with supramolecular surface assemblies. *Nature* **424**, 1029–1031 (2003).
76. Leininger, S., Olenyuk, B. & Stang, P. J. Self-assembly of discrete cyclic nanostructures mediated by transition metals. *Chem. Rev.* **100**, 853–908 (2000).
77. Holiday, B. J. & Mirkin, C. A. Strategies for the construction of supramolecular compounds through coordination chemistry. *Angew. Chem. Int. Edn Engl.* **40**, 2022–2043 (2002).
78. Lingensfelder, M. *et al.* Towards surface-supported supramolecular architectures: tailored coordination assembly of 1,4-benzenedicarboxylate and Fe on Cu(100). *Chem. Eur. J.* **10**, 1913–1919 (2004).
79. Dmitriev, A. *et al.* Design of extended surface-supported chiral metal-organic arrays comprising mononuclear iron centers. *Langmuir* **41**, 4799–4801 (2004).
80. Dmitriev, A., Spillmann, H., Lin, N., Barth, J. V. & Kern, K. Modular assembly of two-dimensional metal-organic coordination networks at a metal surface. *Angew. Chem. Int. Edn Engl.* **41**, 2670–2673 (2003).
81. Stepanow, S. *et al.* Steering molecular organization and host-guest interactions using tailor-made two-dimensional nanoporous coordination systems. *Nature Mater.* **3**, 229–233 (2004).
82. Classen, T. *et al.* Templated growth of metal-organic coordination chains at surfaces. *Angew. Chem. Int. Edn Engl.* (in the press).
83. Messina, P. *et al.* Direct observation of chiral metal-organic complexes assembled on a Cu(100) surface. *J. Am. Chem. Soc.* **124**, 14000–14001 (2002).
84. Spillmann, H. *et al.* Hierarchical assembly of two-dimensional homochiral nanocavity arrays. *J. Am. Chem. Soc.* **125**, 10725–10728 (2003).
85. Yaghi, O. M. *et al.* Reticular synthesis and the design of new materials. *Nature* **423**, 705–714 (2003).
86. Kitagawa, S., Kitaura, R. & Noro, S. Functional coordination polymers. *Angew. Chem. Int. Edn Engl.* **43**, 2334–2375 (2004).
87. Joachim, C., Gimzewski, J. K. & Aviram, A. Electronics using hybrid-molecular and mono-molecular devices. *Nature* **408**, 541–548 (2000).
88. Ouyang, M. & Awschalom, D. D. Coherent spin transfer between molecularly bridged quantum dots. *Science* **301**, 1074–1078 (2003).
89. Niemeyer, C. M. Nanoparticles, proteins and nucleic acids: biotechnology meets materials science. *Angew. Chem. Int. Edn Engl.* **40**, 4128–4158 (2001).
90. Seeman, N. C. & Belcher, A. M. Emulating biology: Building nanostructures from the bottom up. *Proc. Natl Acad. Sci. USA* **99**, 6451–6455 (2002).
91. Sarikaya, M., Tamerler, C., Jen, A. K. Y., Schulten, K. & Baneyx, F. Molecular biomimetics: nanotechnology through biology. *Nature Mater.* **2**, 577–585 (2003).
92. Cleland, A. N. & Roukes, M. L. A nanometre-scale mechanical electrometer. *Nature* **392**, 160–162 (1998).
93. Kind, H. *et al.* Patterned films of nanotubes using microcontact printing of catalysts. *Adv. Mater.* **11**, 1285–1289 (1999).
94. Tans, S. J., Devoret, M. H., Groeneveld, R. J. A. & Dekker, C. Electron-electron correlations in carbon nanotubes. *Nature* **394**, 761–764 (1998).
95. Ross, F. M., Tromp, R. M. & Reuter, M. C. Transition states between pyramids and domes during Ge/Si island growth. *Science* **286**, 1931–1934 (1999).
96. Corso, M. *et al.* Boron nitride nanomesh. *Science* **303**, 217–220 (2004).

Acknowledgements K.K. thanks the many students, postdocs and scientific collaborators who have contributed to the exploration of the atomic world of surfaces and nanostructures. Special thanks go to N. Lin for his enthusiasm in advancing the concepts of supramolecular chemistry at surfaces.

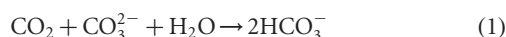
Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.K. (k.kern@fkf.mpg.de).

Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms

James C. Orr¹, Victoria J. Fabry², Olivier Aumont³, Laurent Bopp¹, Scott C. Doney⁴, Richard A. Feely⁵, Anand Gnanadesikan⁶, Nicolas Gruber⁷, Akio Ishida⁸, Fortunat Joos⁹, Robert M. Key¹⁰, Keith Lindsay¹¹, Ernst Maier-Reimer¹², Richard Matear¹³, Patrick Monfray^{1†}, Anne Mouchet¹⁴, Raymond G. Najjar¹⁵, Gian-Kasper Plattner^{7,9}, Keith B. Rodgers^{1,16†}, Christopher L. Sabine⁵, Jorge L. Sarmiento¹⁰, Reiner Schlitzer¹⁷, Richard D. Slater¹⁰, Ian J. Totterdell^{18†}, Marie-France Weirig¹⁷, Yasuhiro Yamanaka⁸ & Andrew Yool¹⁸

Today's surface ocean is saturated with respect to calcium carbonate, but increasing atmospheric carbon dioxide concentrations are reducing ocean pH and carbonate ion concentrations, and thus the level of calcium carbonate saturation. Experimental evidence suggests that if these trends continue, key marine organisms—such as corals and some plankton—will have difficulty maintaining their external calcium carbonate skeletons. Here we use 13 models of the ocean-carbon cycle to assess calcium carbonate saturation under the IS92a 'business-as-usual' scenario for future emissions of anthropogenic carbon dioxide. In our projections, Southern Ocean surface waters will begin to become undersaturated with respect to aragonite, a metastable form of calcium carbonate, by the year 2050. By 2100, this undersaturation could extend throughout the entire Southern Ocean and into the subarctic Pacific Ocean. When live pteropods were exposed to our predicted level of undersaturation during a two-day shipboard experiment, their aragonite shells showed notable dissolution. Our findings indicate that conditions detrimental to high-latitude ecosystems could develop within decades, not centuries as suggested previously.

Ocean uptake of CO₂ will help moderate future climate change, but the associated chemistry, namely hydrolysis of CO₂ in seawater, increases the hydrogen ion concentration [H⁺]. Surface ocean pH is already 0.1 unit lower than preindustrial values. By the end of the century, it will become another 0.3–0.4 units lower^{1,2} under the IS92a scenario, which translates to a 100–150% increase in [H⁺]. Simultaneously, aqueous CO₂ concentrations [CO₂(aq)] will increase and carbonate ion concentrations [CO₃²⁻] will decrease, making it more difficult for marine calcifying organisms to form biogenic calcium carbonate (CaCO₃). Substantial experimental evidence indicates that calcification rates will decrease in low-latitude corals^{3–5}, which form reefs out of aragonite, and in phytoplankton that form their tests (shells) out of calcite^{6,7}, the stable form of CaCO₃. Calcification rates will decline along with [CO₃²⁻] owing to its reaction with increasing concentrations of anthropogenic CO₂ according to the following reaction:



These rates decline even when surface waters remain supersaturated with respect to CaCO₃, a condition that previous studies have predicted will persist for hundreds of years^{4,8,9}.

Recent predictions of future changes in surface ocean pH and carbonate chemistry have primarily focused on global average conditions^{1,2,10} or on low latitude regions⁴, where reef-building corals are abundant. Here we focus on future surface and subsurface changes in high latitude regions where planktonic shelled pteropods are prominent components of the upper-ocean biota in the Southern Ocean, Arctic Ocean and subarctic Pacific Ocean^{11–15}. Recently, it has been suggested that the cold surface waters in such regions will begin to become undersaturated with respect to aragonite only when atmospheric CO₂ reaches 1,200 p.p.m.v., more than four times the preindustrial level (4 × CO₂) of 280 p.p.m.v. (ref. 9). In contrast, our results suggest that some polar and subpolar surface waters will become undersaturated at ~2 × CO₂, probably within the next 50 years.

¹Laboratoire des Sciences du Climat et de l'Environnement, UMR CEA-CNRS, CEA Saclay, F-91191 Gif-sur-Yvette, France. ²Department of Biological Sciences, California State University San Marcos, San Marcos, California 92096-0001, USA. ³Laboratoire d'Océanographie et du Climat: Expérimentations et Approches Numériques (LOCEAN), Centre IRD de Bretagne, F-29280 Plouzané, France. ⁴Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543-1543, USA. ⁵National Oceanic and Atmospheric Administration (NOAA)/Pacific Marine Environmental Laboratory, Seattle, Washington 98115-6349, USA. ⁶NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey 08542, USA. ⁷Institute of Geophysics and Planetary Physics, UCLA, Los Angeles, California 90095-4996, USA. ⁸Frontier Research Center for Global Change, Yokohama 236-0001, Japan. ⁹Climate and Environmental Physics, Physics Institute, University of Bern, CH-3012 Bern, Switzerland. ¹⁰Atmospheric and Oceanic Sciences (AOS) Program, Princeton University, Princeton, New Jersey 08544-0710, USA. ¹¹National Center for Atmospheric Research, Boulder, Colorado 80307-3000, USA. ¹²Max Planck Institut für Meteorologie, D-20146 Hamburg, Germany. ¹³CSIRO Marine Research and Antarctic Climate and Ecosystems CRC, Hobart, Tasmania 7001, Australia. ¹⁴Astrophysics and Geophysics Institute, University of Liege, B-4000 Liege, Belgium. ¹⁵Department of Meteorology, Pennsylvania State University, University Park, Pennsylvania 16802-5013, USA. ¹⁶LOCEAN, Université Pierre et Marie Curie, F-75252 Paris, France. ¹⁷Alfred Wegener Institute for Polar and Marine Research, D-27515 Bremerhaven, Germany. ¹⁸National Oceanography Centre Southampton, Southampton SO14 3ZH, UK. †Present addresses: Laboratoire d'Etudes en Géophysique et Océanographie Spatiales, UMR 5566 CNES-CNRS-IRD-UPS, F-31401 Toulouse, France (P.M.); AOS Program, Princeton University, Princeton, New Jersey 08544-0710, USA (K.B.R.); The Met Office, Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK (I.J.T.).

Changes in carbonate

We have computed modern-day ocean carbonate chemistry from observed alkalinity and dissolved inorganic carbon (DIC), relying on data collected during the CO₂ Survey of the World Ocean Circulation Experiment (WOCE) and the Joint Global Ocean Flux Study (JGOFS). These observations are centred around the year 1994, and have recently been provided as a global-scale, gridded data product GLODAP (ref. 16; see Supplementary Information). Modern-day surface [CO₃²⁻] varies meridionally by more than a factor of two, from average concentrations in the Southern Ocean of 105 μmol kg⁻¹ to average concentrations in tropical waters of 240 μmol kg⁻¹ (Fig. 1). Low [CO₃²⁻] in the Southern Ocean is due to (1) low surface temperatures and CO₂-system thermodynamics, and (2) large amounts of upwelled deep water, which contain high [CO₂(aq)] from organic matter remineralization. These two effects reinforce one another, yielding a high positive correlation of present-day [CO₃²⁻] with temperature (for example, $R^2 = 0.92$ for annual

mean surface maps). Changes in [CO₃²⁻] and [CO₂(aq)] are also inextricably linked to changes in other carbonate chemistry variables (Supplementary Fig. S1).

We also estimated preindustrial [CO₃²⁻] from the same data, after subtracting data-based estimates of anthropogenic DIC (ref. 17) from the modern DIC observations and assuming that preindustrial and modern alkalinity fields were identical (see Supplementary Information). Relative to preindustrial conditions, invasion of anthropogenic CO₂ has already reduced modern surface [CO₃²⁻] by more than 10%, that is, a reduction of 29 μmol kg⁻¹ in the tropics and 18 μmol kg⁻¹ in the Southern Ocean. Nearly identical results were found when, instead of the data-based anthropogenic CO₂ estimates, we used simulated anthropogenic CO₂, namely the median from 13 models that participated in the second phase of the Ocean Carbon-Cycle Model Intercomparison Project, or OCMIP-2 (Fig. 1c).

To quantify future changes in carbonate chemistry, we used simulated DIC from ocean models that were forced by two atmospheric CO₂ scenarios: the Intergovernmental Panel on Climate Change (IPCC) IS92a 'continually increasing' scenario (788 p.p.m.v. in the year 2100) and the IPCC S650 'stabilization' scenario (563 p.p.m.v. in the year 2100) (Fig. 1). Simulated perturbations in DIC relative to 1994 (the GLODAP reference year) were added to the modern DIC data; again, alkalinity was assumed to be constant. To provide a measure of uncertainty, we report model results as the OCMIP median ± 2σ. The median generally outperformed

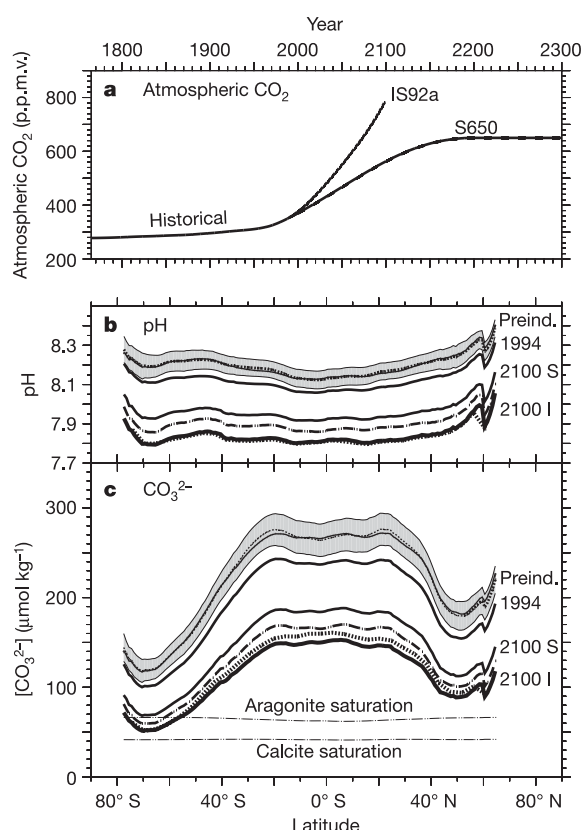


Figure 1 | Increasing atmospheric CO₂ and decreasing surface ocean pH and [CO₃²⁻]. **a**, Atmospheric CO₂ used to force 13 OCMIP models over the industrial period ('Historical') and for two future scenarios: IS92a ('I' in **b** and **c**) and S650 ('S' in **b** and **c**). **b**, **c**, Increases in atmospheric CO₂ lead to reductions in surface ocean pH (**b**) and surface ocean [CO₃²⁻] (**c**). Results are given as global zonal averages for the 1994 data and the preindustrial ('Preind.') ocean. The latter were obtained by subtracting data-based anthropogenic DIC (ref. 17) (solid line in grey-shaded area), as well as by subtracting model-based anthropogenic DIC (OCMIP median, dotted line in grey-shaded area; OCMIP range, grey shading). Future results for the year 2100 come from the 1994 data plus the simulated DIC perturbations for the two scenarios; results are also shown for the year 2300 with S650 (thick dashed line). The small effect of future climate change simulated by the IPSL climate-carbon model is added as a perturbation to IS92a in the year 2100 (thick dotted line); two other climate-carbon models, PIUB-Bern and Commonwealth Scientific and Industrial Research Organisation (CSIRO), show similar results (Fig. 3a). The thin dashed lines indicating the [CO₃²⁻] for sea water in equilibrium with aragonite and calcite are nearly flat, revealing weak temperature sensitivity.

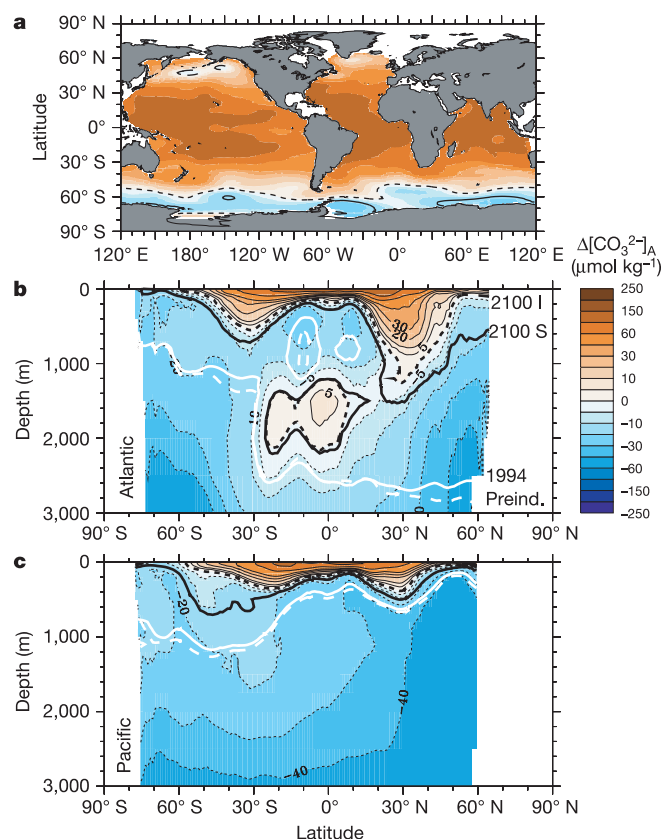


Figure 2 | The aragonite saturation state in the year 2100 as indicated by $\Delta[\text{CO}_3^{2-}]_A$. The $\Delta[\text{CO}_3^{2-}]_A$ is the *in situ* [CO₃²⁻] minus that for aragonite-equilibrated sea water at the same salinity, temperature and pressure. Shown are the OCMIP-2 median concentrations in the year 2100 under scenario IS92a: **a**, surface map; **b**, Atlantic; and **c**, Pacific zonal averages. Thick lines indicate the aragonite saturation horizon in 1765 (Preind.; white dashed line), 1994 (white solid line) and 2100 (black solid line for S650; black dashed line for IS92a). Positive $\Delta[\text{CO}_3^{2-}]_A$ indicates supersaturation; negative $\Delta[\text{CO}_3^{2-}]_A$ indicates undersaturation.

individual models in OCMIP model–data comparison (Supplementary Fig. S2). By the year 2100, as atmospheric CO_2 reaches 788 p.p.m.v. under the IS92a scenario, average tropical surface $[\text{CO}_3^{2-}]$ declines to $149 \pm 14 \mu\text{mol kg}^{-1}$. This is a 45% reduction relative to preindustrial levels, in agreement with previous predictions^{4,8}. In the Southern Ocean (all waters south of 60°S), surface concentrations dip to $55 \pm 5 \mu\text{mol kg}^{-1}$, which is 18% below the threshold where aragonite becomes undersaturated ($66 \mu\text{mol kg}^{-1}$).

These changes extend well below the sea surface. Throughout the Southern Ocean, the entire water column becomes undersaturated with respect to aragonite. During the twenty-first century, under the IS92a scenario, the Southern Ocean's aragonite saturation horizon (the limit between undersaturation and supersaturation) shoals from its present average depth of 730 m (Supplementary Fig. S3) all the way to the surface (Fig. 2). Simultaneously, in a portion of the subarctic Pacific, the aragonite saturation horizon shoals from depths of about 120 m to the surface. In the North Atlantic, surface waters remain saturated with respect to aragonite, but the aragonite saturation horizon shoals dramatically; for example, north of 50°N it shoals from 2,600 m to 115 m. The greater erosion in the North Atlantic is due to deeper penetration and higher concentrations of anthropogenic CO_2 , a tendency that is already evident in present-day data-based estimates^{17,18} and in models^{19,20} (Supplementary Figs S4 and S5). Less pronounced changes were found for the calcite saturation horizon. For example, in the year 2100 the average calcite saturation horizon in the Southern Ocean stays below 2,200 m. Nonetheless, in 2100 surface waters of the Weddell Sea become slightly undersaturated with respect to calcite.

In the more conservative S650 scenario, the atmosphere reaches $2 \times \text{CO}_2$ in the year 2100, 50 years later than with the IS92a scenario. In 2100, Southern Ocean surface waters generally remain slightly supersaturated with respect to aragonite. However, the models also

simulate that the Southern Ocean's average aragonite saturation horizon will have shoaled from 730 m to 60 m, and that the entire water column in the Weddell Sea will have become undersaturated (Fig. 2). In the north, all surface waters remain saturated under the S650 scenario. North of 50°N , the annual average aragonite saturation horizon shoals from 140 m to 70 m in the Pacific, whereas it shoals by 2,000 m to 610 m in the North Atlantic. Therefore, under either scenario the OCMIP models simulated large changes in surface and subsurface $[\text{CO}_3^{2-}]$. Yet these models account for only the direct geochemical effect of increasing atmospheric CO_2 because they were all forced with prescribed modern-day climate conditions.

In addition to this direct geochemical effect, ocean $[\text{CO}_3^{2-}]$ is also altered by climate variability and climate change. To quantify the added effect of future climate change, we analysed results from three atmosphere–ocean climate models that each included an ocean carbon-cycle component (see Supplementary Information). These three models agree that twenty-first century climate change will cause a general increase in surface ocean $[\text{CO}_3^{2-}]$ (Fig. 3), mainly because most surface waters will be warmer. However, the models also agree that the magnitude of this increase in $[\text{CO}_3^{2-}]$ is small, typically counteracting less than 10% of the decrease due to the geochemical effect. High-latitude surface waters show the smallest increases in $[\text{CO}_3^{2-}]$, and even small reductions in some cases. Therefore, our analysis suggests that physical climate change alone will not substantially alter high-latitude surface $[\text{CO}_3^{2-}]$ during the twenty-first century.

Climate also varies seasonally and interannually, whereas our previous focus has been on annual changes. To illustrate how climate variability affects surface $[\text{CO}_3^{2-}]$, we used results from an ocean carbon-cycle model forced with the daily National Centers for Environmental Prediction (NCEP) reanalysis fields²¹ over 1948–2003 (see Supplementary Information). These fields are observationally based and vary on seasonal and interannual timescales. Simulated interannual variability in surface ocean $[\text{CO}_3^{2-}]$ is negligible when compared with the magnitude of the anthropogenic decline (Fig. 3b). Seasonal variability is also negligible except in the high latitudes, where surface $[\text{CO}_3^{2-}]$ varies by about $\pm 15 \mu\text{mol kg}^{-1}$

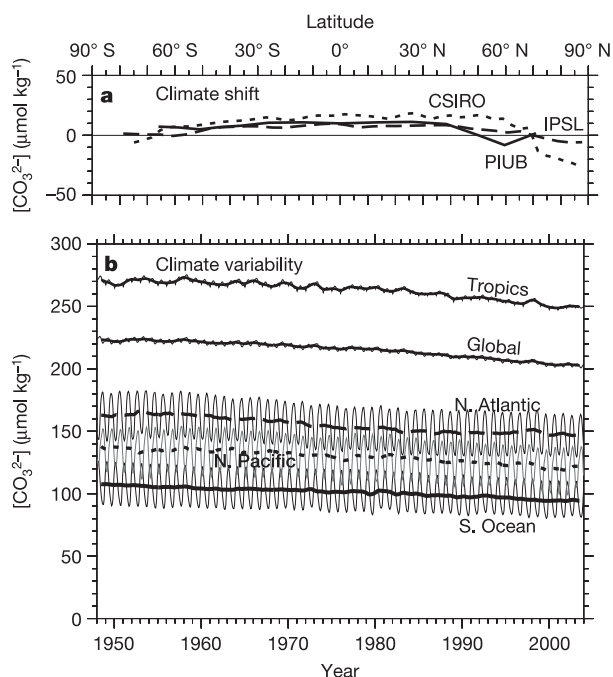


Figure 3 | Climate-induced changes in surface $[\text{CO}_3^{2-}]$. **a**, The twenty-first century shift in zonal mean surface ocean $[\text{CO}_3^{2-}]$ due to climate change alone, from three atmosphere–ocean climate models—CSIRO–Hobart (short dashed line), IPSL–Paris (long dashed line) and PIUB–Bern (solid line)—that each include an ocean carbon-cycle component (see Supplementary Information). **b**, The regional-scale seasonal and interannual variability is simulated by an ocean carbon-cycle model forced with reanalysed climate forcing.

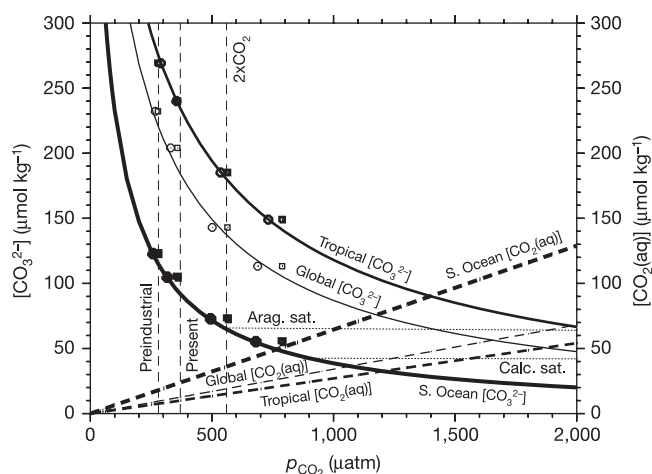


Figure 4 | Key surface carbonate chemistry variables as a function of $p\text{CO}_2$. Shown are both $[\text{CO}_3^{2-}]$ (solid lines) and $[\text{CO}_2(\text{aq})]$ (dashed lines) for average surface waters in the tropical ocean (thick lines), the Southern Ocean (thickest lines) and the global ocean (thin lines). Solid and dashed lines are calculated from the thermodynamic equilibrium approach. For comparison, open symbols are for $[\text{CO}_3^{2-}]$ from our non-equilibrium, model-data approach versus seawater $p\text{CO}_2$ (open circles) and atmospheric $p\text{CO}_2$ (open squares); symbol thickness corresponds with line thickness, which indicates the regions for area-weighted averages. The nearly flat, thin dotted lines indicate the $[\text{CO}_3^{2-}]$ for seawater in equilibrium with aragonite ('Arag. sat.') and calcite ('Calc. sat.').

when averaged over large regions. This is smaller than the twenty-first-century's transient change (for example, $\sim 50 \mu\text{mol kg}^{-1}$ in the Southern Ocean). However, high-latitude surface waters do become substantially less saturated during winter, because of cooling (resulting in higher $[\text{CO}_2(\text{aq})]$) and greater upwelling of DIC-enriched deep water, in agreement with previous observations in the North Pacific²². Thus, high-latitude undersaturation will be first reached during winter.

Our predicted changes may be compared to those found in earlier studies, which focused on surface waters in the tropics⁸ and in the subarctic Pacific^{22,23}. These studies assumed thermodynamic equilibrium between CO_2 in the atmosphere and the surface waters at their *in situ* alkalinity, temperature and salinity. If, in the equilibrium approach, the p_{CO_2} is taken only to represent seawater p_{CO_2} , then the results agree with our non-equilibrium approach when the sets of carbonate chemistry constants are identical (Fig. 4). However, assuming equilibrium with the atmosphere leads to the prediction that future undersaturation will occur too soon (at lower atmospheric CO_2 levels), mainly because the anthropogenic transient in the ocean actually lags that in the atmosphere. For example, with the equilibrium approach, we predict that average surface waters in the Southern Ocean become undersaturated when atmospheric CO_2 is 550 p.p.m.v. (in the year 2050 under IS92a), whereas our non-equilibrium approach, which uses models and data, indicates that undersaturation will occur at 635 p.p.m.v. (in the year 2070). Despite these differences, both approaches indicate that the Southern Ocean surface waters will probably become undersaturated with respect to aragonite during this century. Conversely, both of these approaches disagree with a recent assessment⁹ that used a variant of the standard thermodynamic equilibrium approach, where an incorrect input temperature was used inadvertently.

Uncertainties

The three coupled climate-carbon models show little effect of climate change on surface $[\text{CO}_3^{2-}]$ (compare Fig. 3a to Fig. 1) partly because air-sea CO_2 exchange mostly compensates for the changes in surface DIC caused by changes in marine productivity and circulation. In subsurface waters where such compensation is lacking, these models could under- or over-predict how much $[\text{CO}_3^{2-}]$ will change as a

result of changes in overlying marine productivity. However, the models project a consistent trend, which only worsens the decline in subsurface $[\text{CO}_3^{2-}]$; that is, all coupled climate models predict increased evaporation in the tropics and increased precipitation in the high latitudes²⁴. This leads to greater upper ocean stratification in the high latitudes, which in turn decreases nutrients (but not to zero) and increases light availability (owing to more shallow mixed layers). Thus, at $2 \times \text{CO}_2$ there is a 10% local increase in surface-to-deep export of particulate organic carbon (POC) in the Southern Ocean using the Institut Pierre Simon Laplace (IPSL)-Paris model²⁵. Subsequent remineralization of this exported POC within the thermocline would increase DIC, which would only exacerbate the decrease in high-latitude subsurface $[\text{CO}_3^{2-}]$. For the twenty-first century, these uncertainties appear small next to the anthropogenic DIC invasion (see Supplementary Information).

The largest uncertainty by far, and the only means to limit the future decline in ocean $[\text{CO}_3^{2-}]$, is the atmospheric CO_2 trajectory. To better characterize uncertainty due to CO_2 emissions, we compared the six illustrative IPCC Special Reports on Emission Scenarios (SRES) in the reduced complexity, Physics Institute University of Bern (PIUB)-Bern model. Under the moderate SRES B2 scenario, average Southern Ocean surface waters in that model become undersaturated with respect to aragonite when atmospheric CO_2 reaches 600 p.p.m.v. in the year 2100 (Fig. 5). For the three higher-emission SRES scenarios (A1FI, A2 and A1B), these waters become undersaturated sooner (between the years 2058 and 2073); for the two lower-emission scenarios (A1T and B1), these waters remain slightly supersaturated in 2100. Thus, if atmospheric CO_2 rises above 600 p.p.m.v., most Southern Ocean surface waters will become undersaturated with respect to aragonite. Yet, even below this level, the Southern Ocean's aragonite saturation horizon will shoal substantially (Fig. 2). For a given atmospheric CO_2 scenario, predicted changes in surface ocean $[\text{CO}_3^{2-}]$ are much more certain than the related changes in climate. The latter depend not only on the model response to CO_2 forcing, but also on poorly constrained physical processes, such as those associated with clouds.

Ocean CO_2 uptake

With higher levels of anthropogenic CO_2 and lower surface $[\text{CO}_3^{2-}]$, the change in surface ocean DIC per unit change in atmospheric CO_2 ($\mu\text{mol kg}^{-1}$ per p.p.m.v.) will be about 60% lower in the year 2100 (under IS92a) than it is today. Simultaneously, the $\text{CO}_3^{2-}/\text{CO}_2(\text{aq})$ ratio will decrease from 4:1 to 1:1 in the Southern Ocean (Fig. 4). These decreases are due to the well-understood anthropogenic reduction in buffer capacity²⁶, already accounted for in ocean carbon-cycle models.

On the other hand, reduced export of CaCO_3 from the high latitudes would increase surface $[\text{CO}_3^{2-}]$, thereby increasing ocean CO_2 uptake and decreasing atmospheric CO_2 . Owing to this effect, ocean CO_2 uptake could increase by 6–13 petagrams (Pg) C over the twenty-first century, based on one recent model study²⁷ that incorporated an empirical, CO_2 -dependant relationship for calcification⁷. Rates of calcification could decline even further, to zero, if waters actually became undersaturated with respect to both aragonite and calcite. We estimate that the total shutdown of high-latitude aragonite production would lead to, at most, a $0.25 \text{ Pg C yr}^{-1}$ increase in ocean CO_2 uptake, assuming that 1 Pg C yr^{-1} of CaCO_3 is exported globally²⁸, that up to half of that is aragonite^{9,29}, and that perhaps half of all aragonite is exported from the high latitudes. The actual increase in ocean CO_2 uptake could be much lower because the aragonite fraction of the CaCO_3 may be only 0.1 based on low-latitude sediment traps³⁰, and the latitudinal distribution of aragonite export is uncertain. Thus, increased CO_2 uptake from reduced export of aragonite will provide little compensation for decreases in ocean CO_2 uptake due to reductions in buffer capacity. Of greater concern are potential biological impacts due to future undersaturation.

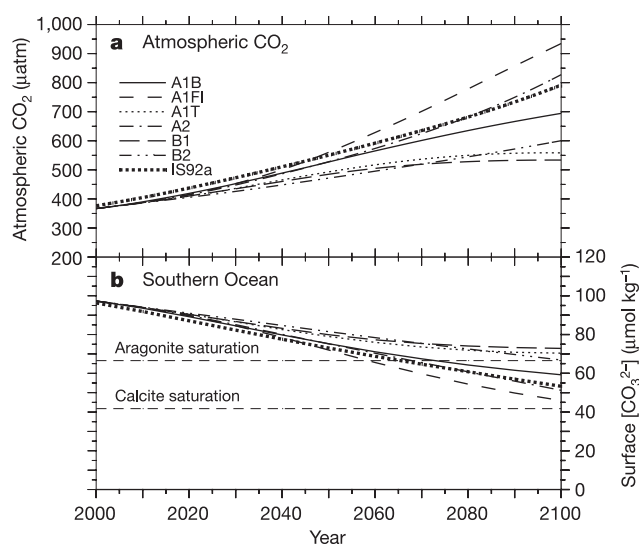


Figure 5 | Average surface $[\text{CO}_3^{2-}]$ in the Southern Ocean under various scenarios. Time series of average surface $[\text{CO}_3^{2-}]$ in the Southern Ocean for the PIUB-Bern reduced complexity model (see Fig. 3 and Supplementary Information) under the six illustrative IPCC SRES scenarios. The results for the SRES scenarios A1T and A2 are similar to those for the non-SRES scenarios S650 and IS92a, respectively.

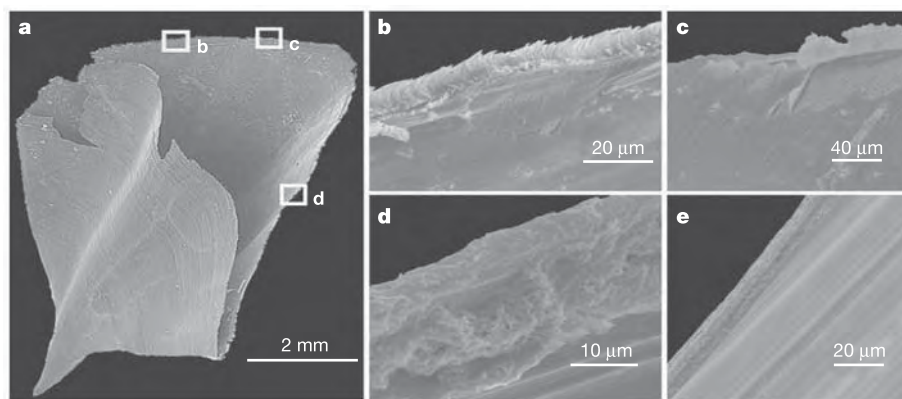


Figure 6 | Shell dissolution in a live pteropod. **a–d**, Shell from a live pteropod, *Clio pyramidata*, collected from the subarctic Pacific and kept in water undersaturated with respect to aragonite for 48 h. The whole shell (**a**) has superimposed white rectangles that indicate three magnified areas: the shell surface (**b**), which reveals etch pits from dissolution and resulting

exposure of aragonitic rods; the prismatic layer (**c**), which has begun to peel back, increasing the surface area over which dissolution occurs; and the aperture region (**d**), which reveals advanced shell dissolution when compared to a typical *C. pyramidata* shell not exposed to undersaturated conditions (**e**).

Biological impacts

The changes in seawater chemistry that we project to occur during this century could have severe consequences for calcifying organisms, particularly shelled pteropods: the major planktonic producers of aragonite. Pteropod population densities are high in polar and subpolar waters. Yet only five species typically occur in such cold water regions and, of these, only one or two species are common at the highest latitudes³¹. High-latitude pteropods have one or two generations per year^{12,15,32}, form integral components of food webs, and are typically found in the upper 300 m where they may reach densities of hundreds to thousands of individuals per m³ (refs 11, 13–15). In the Ross Sea, for example, the prominent subpolar–polar pteropod *Limacina helicina* sometimes replaces krill as the dominant zooplankton, and is considered an overall indicator of ecosystem health³³. In the strongly seasonal high latitudes, sedimentation pulses of pteropods frequently occur just after summer^{15,34}. In the Ross Sea, pteropods account for the majority of the annual export flux of both carbonate and organic carbon^{34,35}. South of the Antarctic Polar Front, pteropods also dominate the export flux of CaCO₃ (ref. 36).

Pteropods may be unable to maintain shells in waters that are undersaturated with respect to aragonite. Data from sediment traps indicate that empty pteropod shells exhibit pitting and partial dissolution as soon as they fall below the aragonite saturation horizon^{22,36,37}. *In vitro* measurements confirm such rapid pteropod shell dissolution rates³⁸. New experimental evidence suggests that even the shells of live pteropods dissolve rapidly once surface waters become undersaturated with respect to aragonite⁹. Here we show that when the live subarctic pteropod *Clio pyramidata* is subjected to a level of undersaturation similar to what we predict for Southern Ocean surface waters in the year 2100 under IS92a, a marked dissolution occurs at the growing edge of the shell aperture within 48 h (Fig. 6). Etch pits formed on the shell surface at the apertural margin (which is typically ~7-μm-thick) as the <1-μm exterior (prismatic layer) peeled back (Fig. 6c), exposing the underlying aragonitic rods to dissolution. Fourteen individuals were tested. All of them showed similar dissolution along their growing edge, even though they all remained alive. If *C. pyramidata* cannot grow its protective shell, we would not expect it to survive in waters that become undersaturated with respect to aragonite.

If the response of other high-latitude pteropod species to aragonite undersaturation is similar to that of *C. pyramidata*, we hypothesize that these pteropods will not be able to adapt quickly enough to live in the undersaturated conditions that will occur over much of the high-latitude surface ocean during the twenty-first century. Their

distributional ranges would then be reduced both within the water column, disrupting vertical migration patterns, and latitudinally, imposing a shift towards lower-latitude surface waters that remain supersaturated with respect to aragonite. At present, we do not know if pteropod species endemic to polar regions could disappear altogether, or if they can make the transition to live in warmer, carbonate-rich waters at lower latitudes under a different ecosystem. If pteropods are excluded from polar and subpolar regions, their predators will be affected immediately. For instance, gymnosomes are zooplankton that feed exclusively on shelled pteropods^{33,39}. Pteropods also contribute to the diet of diverse carnivorous zooplankton, myctophid and nototheniid fishes^{40–42}, North Pacific salmon^{43,44}, mackerel, herring, cod and baleen whales⁴⁵.

Surface dwelling calcitic plankton, such as foraminifera and coccolithophorids, may fare better in the short term. However, the beginnings of high-latitude calcite undersaturation will only lag that for aragonite by 50–100 years. The diverse benthic calcareous organisms in high-latitude regions may also be threatened, including cold-water corals which provide essential fish habitat⁴⁶. Cold-water corals seem much less abundant in the North Pacific than in the North Atlantic⁴⁶, where the aragonite saturation horizon is much deeper (Fig. 2). Moreover, some important taxa in Arctic and Antarctic benthic communities secrete magnesian calcite, which can be more soluble than aragonite. These include gorgonians⁴⁶, coralline red algae and echinoderms (sea urchins)⁴⁷. At 2 × CO₂, juvenile echinoderms stopped growing and produced more brittle and fragile exoskeletons in a subtropical six-month manipulative experiment⁴⁸. However, the responses of high-latitude calcifiers to reduced [CO₃²⁻] have generally not been studied. Yet experimental evidence from many lower-latitude, shallow-dwelling calcifiers reveals a reduced ability to calcify with a decreasing carbonate saturation state⁹. Given that at 2 × CO₂, calcification rates in some shallow-dwelling calcareous organisms may decline by up to 50% (ref. 9), some calcifiers could have difficulty surviving long enough even to experience undersaturation. Certainly, they have not experienced undersaturation for at least the last 400,000 years⁴⁹, and probably much longer⁵⁰.

Changes in high-latitude seawater chemistry that will occur by the end of the century could well alter the structure and biodiversity of polar ecosystems, with impacts on multiple trophic levels. Assessing these impacts is impeded by the scarcity of relevant data.

Received 15 June; accepted 29 July 2005.

1. Haugan, P. M. & Drange, H. Effects of CO₂ on the ocean environment. *Energy Convers. Mgmt* 37, 1019–1022 (1996).

2. Brewer, P. G. Ocean chemistry of the fossil fuel CO₂ signal: the haline signal of "business as usual". *Geophys. Res. Lett.* **24**, 1367–1369 (1997).
3. Gattuso, J.-P., Frankignoulle, M., Bourge, I., Romaine, S. & Buddemeier, R. W. Effect of calcium carbonate saturation of seawater on coral calcification. *Glob. Planet. Change* **18**, 37–46 (1998).
4. Kleypas, J. A. *et al.* Geochemical consequences of increased atmospheric carbon dioxide on coral reefs. *Science* **284**, 118–120 (1999).
5. Langdon, C. *et al.* Effect of elevated CO₂ on the community metabolism of an experimental coral reef. *Glob. Biogeochem. Cycles* **17**, 1011, doi:10.1029/2002GB001941 (2003).
6. Riebesell, U. *et al.* Reduced calcification of marine plankton in response to increased atmospheric CO₂. *Nature* **407**, 364–367 (2000).
7. Zondervan, I., Zeebe, R., Rost, B. & Riebesell, U. Decreasing marine biogenic calcification: A negative feedback on rising atmospheric pCO₂. *Glob. Biogeochem. Cycles* **15**, 507–516 (2001).
8. Broecker, W. S. & Peng, T.-H. Fate of fossil fuel carbon dioxide and the global carbon budget. *Science* **206**, 409–418 (1979).
9. Feely, R. A. *et al.* The impact of anthropogenic CO₂ on the CaCO₃ system in the oceans. *Science* **305**, 362–366 (2004).
10. Caldeira, K. & Wickett, M. E. Anthropogenic carbon and ocean pH. *Nature* **425**, 365 (2003).
11. Urban-Rich, J., Dagg, M. & Peterson, J. Copepod grazing on phytoplankton in the Pacific sector of the Antarctic Polar Front. *Deep-Sea Res. II* **48**, 4223–4246 (2001).
12. Kobayashi, H. A. Growth cycle and related vertical distribution of the thecosomatous pteropod *Spiratella "Limacina" helicina* in the central Arctic Ocean. *Mar. Biol.* **26**, 295–301 (1974).
13. Pakhomov, E. A., Verheye, H. M., Atkinson, A., Laubscher, R. K. & Taunton-Clark, J. Structure and grazing impact of the mesozooplankton community during late summer 1994 near South Georgia, Antarctica. *Polar Biol.* **18**, 180–192 (1997).
14. Fabry, V. J. Aragonite production by pteropod molluscs in the subarctic Pacific. *Deep-Sea Res. I* **36**, 1735–1751 (1989).
15. Bathmann, U., Noji, T. T. & von Bodungen, B. Sedimentation of pteropods in the Norwegian Sea in autumn. *Deep-Sea Res.* **38**, 1341–1360 (1991).
16. Key, R. M. *et al.* A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP). *Glob. Biogeochem. Cycles* **18**, 4031, doi:10.1029/2004GB002247 (2004).
17. Sabine, C. L. *et al.* The ocean sink for anthropogenic CO₂. *Science* **305**, 367–370 (2004).
18. Gruber, N. Anthropogenic CO₂ in the Atlantic Ocean. *Glob. Biogeochem. Cycles* **12**, 165–191 (1998).
19. Sarmiento, J. L., Orr, J. C. & Siegenthaler, U. A perturbation simulation of CO₂ uptake in an ocean general circulation model. *J. Geophys. Res.* **97**, 3621–3645 (1992).
20. Orr, J. C. *et al.* Estimates of anthropogenic carbon uptake from four three-dimensional global ocean models. *Glob. Biogeochem. Cycles* **15**, 43–60 (2001).
21. Kalnay, E. *et al.* The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**, 437–471 (1996).
22. Feely, R. A. *et al.* Winter–summer variations of calcite and aragonite saturation in the northeast Pacific. *Mar. Chem.* **25**, 227–241 (1988).
23. Feely, R. A., Byrne, R. H., Betzer, P. R., Gendron, J. F. & Acker, J. G. Factors influencing the degree of saturation of the surface and intermediate waters of the North Pacific Ocean with respect to aragonite. *J. Geophys. Res.* **89**, 10631–10640 (1984).
24. Sarmiento, J. L. *et al.* Response of ocean ecosystems to climate warming. *Glob. Biogeochem. Cycles* **18**, 3003, doi:10.1029/2003GB002134 (2004).
25. Bopp, L. *et al.* Potential impact of climate change of marine export production. *Glob. Biogeochem. Cycles* **15**, 81–99 (2001).
26. Sarmiento, J. L., Le Quéré, C. & Pacala, S. Limiting future atmospheric carbon dioxide. *Glob. Biogeochem. Cycles* **9**, 121–137 (1995).
27. Heinze, C. Simulating oceanic CaCO₃ export production in the greenhouse. *Geophys. Res. Lett.* **31**, L16308, doi:10.1029/2004GL020613 (2004).
28. Iglesias-Rodriguez, M. D. *et al.* Representing key phytoplankton functional groups in ocean carbon cycle models: Coccolithophorids. *Glob. Biogeochem. Cycles* **16**, 1100, doi:10.1029/2001GB001454 (2002).
29. Berner, R. A. in *The Fate of Fossil Fuel CO₂ in the Oceans* (eds Andersen, N. R. & Malahoff, A.) 243–260 (Plenum, New York, 1977).
30. Fabry, V. J. Shell growth rates of pteropod and heteropod molluscs and aragonite production in the open ocean: Implications for the marine carbonate system. *J. Mar. Res.* **48**, 209–222 (1990).
31. Bé, A. W. H. & Gilmer, R. W. *Oceanic Micropaleontology* Vol. 1 (ed. Ramsey, A.) 733–808 (Academic, London, 1977).
32. Dadon, J. R. & de Cidre, L. L. The reproductive cycle of the Thecosomatous pteropod *Limacina retroversa* in the western South Atlantic. *Mar. Biol.* **114**, 439–442 (1992).
33. Seibel, B. A. & Dierssen, H. M. Cascading trophic impacts of reduced biomass in the Ross Sea, Antarctica: Just the tip of the iceberg? *Biol. Bull.* **205**, 93–97 (2003).
34. Accornero, A., Manno, C., Esposito, F. & Gambi, M. C. The vertical flux of particulate matter in the polyna of Terra Nova Bay. Part II. Biological components. *Antarct. Sci.* **15**, 175–188 (2003).
35. Collier, R., Dymond, J., Susumu Honjo, S. M., Francois, R. & Dunbar, R. The vertical flux of biogenic and lithogenic material in the Ross Sea: moored sediment trap observations 1996–1998. *Deep-Sea Res. II* **47**, 3491–3520 (2000).
36. Honjo, S., Francois, R., Manganini, S., Dymond, J. & Collier, R. Particle fluxes to the interior of the Southern Ocean in the western Pacific sector along 170° W. *Deep-Sea Res. II* **47**, 3521–3548 (2000).
37. Betzer, P. R., Byrne, R., Acker, J., Lewis, C. S. & Jolley, R. R. The oceanic carbonate system: a reassessment of biogenic controls. *Science* **226**, 1074–1077 (1984).
38. Byrne, R. H., Acker, J. G., Betzer, P. R., Feely, R. A. & Cates, M. H. Water column dissolution of aragonite in the Pacific Ocean. *Nature* **312**, 321–326 (1984).
39. Lalli, C. M. Structure and function of the buccal apparatus of *Clione limacina* (Phipps) with a review of feeding in gymnosomatous pteropods. *J. Exp. Mar. Biol. Ecol.* **4**, 101–118 (1970).
40. Foster, B. A. & Montgomery, J. C. Planktivory in benthic nototheniid fish in McMurdo Sound, Antarctica. *Environ. Biol. Fish.* **36**, 313–318 (1993).
41. Pakhomov, E., Perissinotto, A. & McQuaid, C. D. Prey composition and daily rations of myctophid fishes in the Southern Ocean. *Mar. Ecol. Prog. Ser.* **134**, 1–14 (1996).
42. La Mesa, M., Vacchi, M. & Sertorio, T. Z. Feeding plasticity of *Trematomus newnesi* (Pisces, Nototheniidae) in Terra Nova Bay, Ross Sea, in relation to environmental conditions. *Polar Biol.* **23**, 38–45 (2000).
43. Willette, T. M. *et al.* Ecological processes influencing mortality of juvenile pink salmon (*Oncorhynchus gorbuscha*) in Prince William Sound, Alaska. *Fish. Oceanogr.* **10**, 14–41 (2001).
44. Boldt, J. L. & Halderson, L. J. Seasonal and geographical variation in juvenile pink salmon diets in the Northern Gulf of Alaska and Prince William Sound. *Trans. Am. Fisheries Soc.* **132**, 1035–1052 (2003).
45. Lalli, C. M. & Gilmer, R. *Pelagic Snails* (Stanford Univ. Press, Stanford, 1989).
46. Freiwald, A., Fosså, J. H., Grehan, A., Koslow, T. & Roberts, J. M. *Cold-water Coral Reefs: Out of Sight—No Longer Out of Mind* (No. 22 in Biodiversity Series, UNEP-WCMC, Cambridge, UK, 2004).
47. Dayton, P. K. in *Polar Oceanography, Part B: Chemistry, Biology and Geology* (ed. Smith, W. O.) 631–685 (Academic, San Diego, 1990).
48. Shirayama, Y. & Thornton, H. Effect of increased atmospheric CO₂ on shallow-water marine benthos. *J. Geophys. Res.* **110**, C09S09, doi:10.1029/2004JC002561 (2005).
49. Petit, J. R. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436 (1999).
50. Pearson, P. N. & Palmer, M. R. Middle Eocene seawater pH and atmospheric carbon dioxide concentrations. *Science* **284**, 1824–1826 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Gehlen for discussions, and J.-M. Epitalon, P. Brockmann and the Ferret developers for help with analysis. All but the climate simulations were made as part of the OCMIP project, which was launched in 1995 by the Global Analysis, Integration and Modelling (GAIM) Task Force of the International Geosphere-Biosphere Programme (IGBP) with funding from NASA (National Aeronautics and Space Administration). OCMIP-2 was supported by the European Union Global Ocean Storage of Anthropogenic Carbon (EU GOSAC) project and the United States JGOFS Synthesis and Modeling Project funded through NASA. The interannual simulation was supported by the EU Northern Ocean Carbon Exchange Study (NOCES) project, which is part of OCMIP-3.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.C.O. (orr@cea.fr).

The effect of advection on the nutrient reservoir in the North Atlantic subtropical gyre

Jaime B. Palter¹, M. Susan Lozier¹ & Richard T. Barber²

Though critically important in sustaining the ocean's biological pump, the cycling of nutrients in the subtropical gyres is poorly understood. The supply of nutrients to the sunlit surface layer of the ocean has traditionally been attributed solely to vertical processes. However, horizontal advection may also be important in establishing the availability of nutrients. Here we show that the production and advection of North Atlantic Subtropical Mode Water introduces spatial and temporal variability in the subsurface nutrient reservoir beneath the North Atlantic subtropical gyre. As the mode water is formed, its nutrients are depleted by biological utilization. When the depleted water mass is exported to the gyre, it injects a wedge of low-nutrient water into the upper layers of the ocean. Contrary to intuition, cold winters that promote deep convective mixing and vigorous mode water formation may diminish downstream primary productivity by altering the subsurface delivery of nutrients.

Wind-driven upwelling¹, diapycnal diffusion², wintertime convection^{3–5} and, more recently, eddy heaving^{6,7} have all been examined for their contribution to the upward flux of nutrients in the global ocean. It is generally assumed that the variability in these vertical processes translates into variability of primary productivity. This assumption relies on the premise that an adequate reservoir of nutrients resides below the euphotic zone—this is a possibility only if the remineralization of organic matter at depth occurs much more quickly than the physical processes that move nutrients upwards. In this one-dimensional view, the downward flux of organic material from the surface ocean is balanced locally by the upward flux of nutrients from the thermocline, as discussed in ref. 8. Though conceptually appealing in its simplicity, this view not only neglects the horizontal advection of phytoplankton into or out of a locale, but also neglects the lateral processes that deliver nutrients to the subsurface. The production and advection of Subtropical Mode Water (STMW) is one such lateral process that affects the nutrient reservoir of the North Atlantic subtropical gyre.

Nutrient depletion of the STMW during formation

STMW is formed by convection each winter in an east–west band at the northern edge of the subtropical gyre, just south of the Gulf Stream^{9,10} (Fig. 1). As the water mass leaves the formation region, it is capped by warming surface waters in the oncoming spring, and then subducted beneath even warmer water as it travels to the south, transiting the subtropical gyre¹¹. The nutrient concentrations in the STMW change rapidly during formation. Nutrient concentrations in the STMW formation region are consistently negligible to the base of the euphotic zone, located at roughly 120 m depth, whereas wintertime convective mixing typically reaches between 200 and 400 m depth. Thus, convective mixing entrains nutricline fluid into the mixed layer, where it is combined with zero-nutrient euphotic zone water. The resulting nutrient concentrations reflect the properties of both the euphotic zone and the underlying nutricline, as well as any

ongoing biological utilization. Such utilization continuously competes with the entrainment flux of nutrients in setting the mixed layer concentrations during formation, as made apparent by the timing of the seasonal chlorophyll bloom. The initiation of the seasonal

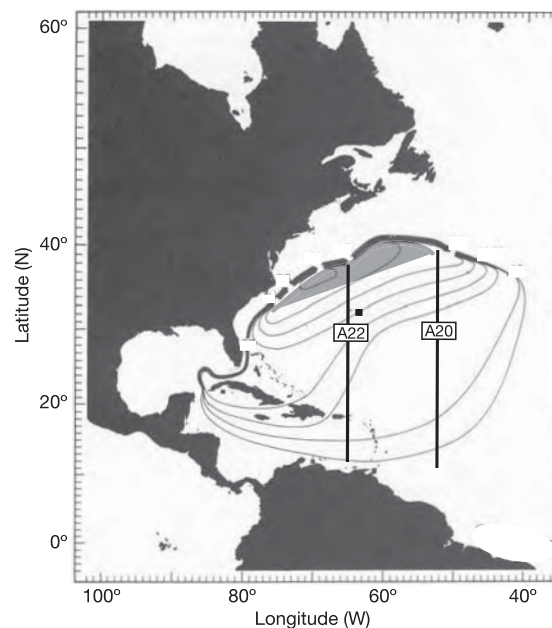


Figure 1 | Schematic representation of the mean circulation of the warm water (>17°C) of the North Atlantic. The filled square shows the approximate location of the BATS and Hydrostation S time series. Thick vertical lines show the WOCE repeat cruise tracks (A22 and A20) used for this study. Grey shading shows the approximate location of the STMW formation region. The solid grey lines represent streamlines with the greatest transport. Modified from ref. 11.

¹Division of Earth and Ocean Science, Nicholas School of the Environment and Earth Science, Duke University, Durham, North Carolina 27708, USA. ²Division of Coastal Systems Science and Policy, Nicholas School of the Environment and Earth Science, Duke University Marine Laboratory, 135 Duke Marine Lab Road, Beaufort, North Carolina 28516, USA.

chlorophyll bloom in the region of 30–40° N occurs in January¹², with the maximum chlorophyll concentrations occurring in March and April. The coincidence of the winter bloom initiation with the annual solar insolation minimum precludes light as the major limiting factor in phytoplankton growth, and instead points towards nutrient limitation. The timing of the winter bloom suggests that phytoplankton are actively depleting nutrients during convective events that entrain sub-euphotic-zone water into the mixed layer. The result is that nitrate and phosphate concentrations in STMW at the time of subduction are 10–20% of those found in the upstream source waters¹³ (see Methods), which may derive in part from eroded STMW that has joined the Gulf Stream after transiting the gyre.

Downstream effect of STMW

The physical signature of STMW is a subsurface thermostad (a thick layer of nearly uniform temperature) centred roughly at 18 °C (Fig. 2a, c). From the representative hydrographic profiles in the subtropical gyre (Fig. 2), it is clear that the presence and thickness of the STMW varies in space and time, as does the attendant nutrient concentration below the euphotic zone. Here we use nitrate specifically for illustration, though the patterns are identical for phosphate, as the two are highly correlated for all data used in this analysis. For example, near Bermuda, downstream of the STMW formation region (Fig. 1), the mode water was roughly 400 m thick in July 1960, with nitrate concentrations below 2.5 mmol m⁻³ to a depth of 500 m (Fig. 2a). By contrast, in July 1989 near the same site, no more than 200 m of an eroded thermostad exists, the concentration of nitrate at 500 m depth is 7.1 mmol m⁻³, and the integrated nitrate from the surface to 500 m is almost twice that during the same month in 1960 (Fig. 2b).

This linkage between the presence of STMW and low nutrient concentration extends to the spatial domain as well, as illustrated by two profiles taken on the same World Ocean Circulation Experiment (WOCE) survey in July 1997. The presence of STMW is clear inside the region of subtropical recirculation (Fig. 2c), but not in a more southerly profile from outside the realm where STMW resides (Fig. 2d, note the absence of a thermostad near 18 °C). The change in the nutricline depth across the subtropical gyre can be explained by the presence of the wedge of STMW, which is thickest near its source and thins southward. The underlying thermocline and nutricline move downwards to accommodate that wedge. In addition, the thermostad itself is relatively low in nutrients, so above the depressed

nutricline the inserted STMW is a low concentration 'nutrirstad'. Thus, it appears that in regions and years with a strong STMW signature, low-nitrate waters reside beneath the euphotic zone. In regions and years lacking the characteristic STMW thermostad, the nutricline, no longer depressed, is a steep and nearly linear gradient between the base of the euphotic zone and the remineralized nutrients at depth.

Spatial variability of the nutrient reservoir

The effect of the STMW on the subsurface nutrient gradient is more thoroughly demonstrated by examining a WOCE meridional section across the subtropical North Atlantic (Fig. 3). The STMW is apparent as a wedge of nearly uniform density water between the depths of 150 and 400 m, to the north of 20° N (Fig. 3a). As evidenced from an inspection of the nitrate cross-section, the presence of the STMW coincides with the deepening of the nutricline (Fig. 3b). Because mode water is convectively formed, it is characterized by a low vertical density gradient and, thus, a local minimum in the absolute value of potential vorticity (PV) (Fig. 3c)¹⁰. PV is a convenient tracer of the water mass, as it reflects the local density gradient and tends to be preserved away from the formation region as it has no internal sources or sinks. In the WOCE nutrient sections, a noticeable front in nitrate at the edge of the PV minimum (at roughly 20° N) provides evidence that STMW has a much lower nutrient concentration than surrounding water at the same density (Fig. 3d). Along a single density surface, nitrate concentrations within the STMW are roughly half that outside the zone of STMW recirculation¹¹. These richer waters are further from their ventilation sources, and thus older.

The low-nutrient signature of the STMW, set during ventilation, can be seen more than 2,000 km to the south of the source region. This depletion persists despite the ongoing remineralization of nutrients at depth that acts to annihilate the low nitrate concentration of the water mass. The depletion also persists in the presence of vertical fluxes that act to weaken the nutrirstad. To understand these competing processes, the advective, remineralization and diffusive timescales have been estimated from a scale analysis of the nitrate conservation equation (see Methods). A relatively large timescale for turbulent diffusion suggests that this process is minimally important in setting the nutrient concentrations within the nutrient reservoir. However, the ratio of the advective timescale to the remineralization timescale is of the order of one within the subtropical gyre, suggesting an important competition between these two processes. Indeed, the effect of the persistent remineralization is manifested by the vertical nutrient gradient within the low-PV water mass (Fig. 3d, e). A uniform nutrient concentration within the STMW is expected at formation, yet remineralization begins re-establishing a vertical gradient once the water mass is subducted. At the PV minimum, we estimate a nitrate remineralization rate of 0.53 ± 0.14 mmol m⁻³ yr⁻¹, too slow to completely restore STMW nitrate to its pre-formation concentration in the time it takes to transit the gyre. The relationship between STMW and nutrients implied from an inspection of the WOCE meridional sections is quantitatively assessed in Fig. 4, where PV and its corresponding nitrate concentration are plotted. Here, we are making the explicit assumption that PV is a proxy for the age of the water mass. This is an appropriate assumption given that the low-PV signature of STMW is slowly eroded over time by diffusive processes. Indeed, PV and chlorofluorocarbon (CFC) age are significantly correlated within the STMW ($r = 0.74$, Supplementary Fig. S1) along the WOCE sections for which CFC age is available (see Methods). Low-PV waters, indicative of recent ventilation, have lower nitrate concentrations than waters of the same density with higher PV. Using the WOCE data displayed in Fig. 4, PV is significantly correlated with nitrate ($r = 0.86$).

Temporal variability of the nutrient reservoir

Because properties of the mode water do not reflect just the response

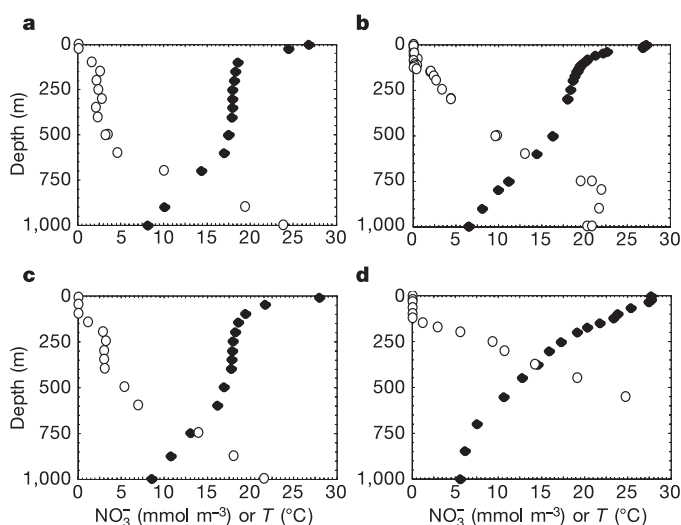


Figure 2 | Nitrate (open symbols) and temperature (filled symbols) as a function of depth. Data are shown for: **a**, Hydrostation S (32.17° N, 64.50° W) in July 1960; **b**, BATS (31.92° N, 64.17° W) in July 1989; **c**, 30.2° N 52.3° W, from WOCE section A20 in July 1997; and **d**, 16.2° N 52.3° W, from the same WOCE section A20 in July 1997.

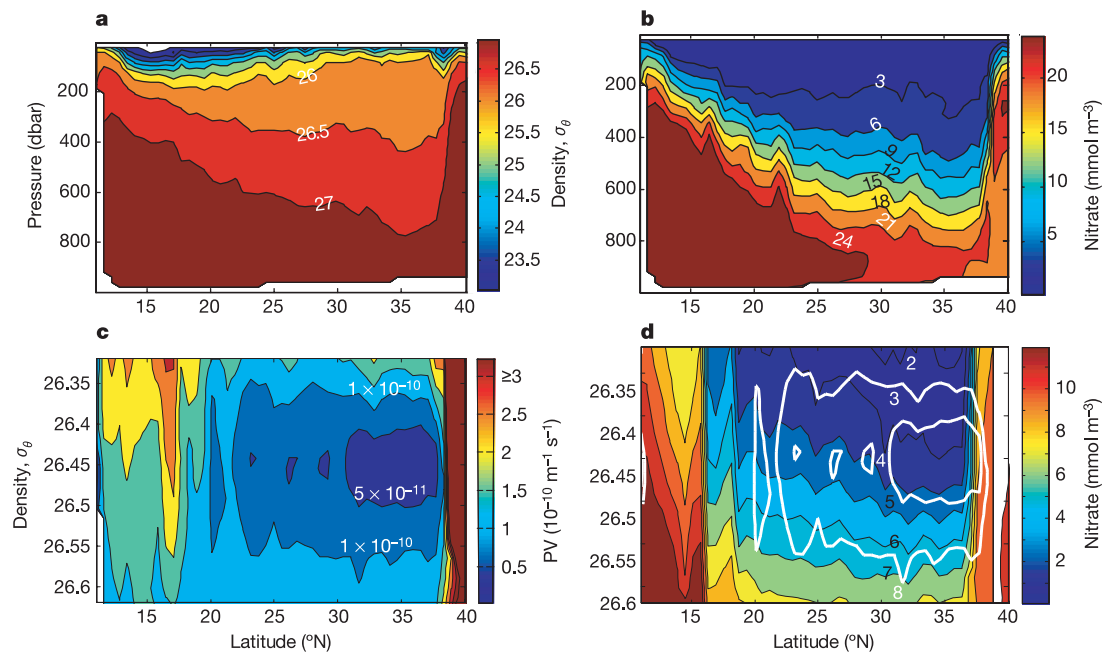


Figure 3 | Properties of WOCE section A22 in August 1997. a, Potential density as a function of pressure. **b,** Nitrate as a function of pressure. **c,** Potential vorticity (PV) as a function of potential density; the low-PV waters ($\leq 1 \times 10^{-10} \text{ m}^{-1} \text{ s}^{-1}$, shaded blue) are considered the core of the

STMW. **d,** Nitrate as a function of potential density. The white contour lines in **d** represent $PV = -0.5 \times 10^{-10}$ and $-1 \times 10^{-10} \text{ m}^{-1} \text{ s}^{-1}$. PV was calculated using $f/\sigma_\theta (\partial \sigma_\theta / \partial z)$, where f is the Coriolis parameter, σ_θ the reference density, and $\partial \sigma_\theta / \partial z$ the vertical density gradient.

to the current year's forcing, but the accumulated effects of several years^{14,15}, it is expected that extended periods of sustained deep mixing introduce interdecadal variability to the STMW nutrient reservoir downstream of the formation region. To test this expectation, two time series of nutrient data collected near Bermuda, one during the Hydrostation S programme from 1958–63 and another from the Bermuda Atlantic Time Series (BATS) programme from 1988–present, were compared. Throughout the years of the Hydrostation S observations, the North Atlantic Oscillation (NAO) was in a predominantly negative phase and relatively cold conditions in the subtropics produced dense, thick STMW. In contrast, the predominantly positive NAO regime during the BATS era caused sluggish mixing and low STMW production^{15,16}. From our analysis, the mean nitrate concentration within the STMW is 25% higher in the positive NAO years than in the negative NAO years, a significant difference at the 1% level. Importantly, the water-column-integrated nitrate, from the surface to the top of the permanent pycnocline (nominally 400 m), is also slightly higher in the positive NAO years. This is opposite to the expectation that deeper mixing should be associated with greater nutrient availability. Instead, the time series data provide compelling evidence that low-nutrient STMW is exported to the subtropical gyre beneath the euphotic zone in periods of cold winters and intense convective mixing. As a result, the available nutrient reservoir is reduced. The correlation between PV and nitrate for the time series data (displayed in Fig. 4, $r = 0.56$) indicates that the mode water present at Bermuda is most depleted in nitrate when it is most recently formed. Thus, whatever the process that moves subsurface water upwards, the delivery of nutrients will be damped in years when low-nutrient STMW occupies the subsurface nutrient reservoir.

Primary productivity from 1958–60 at Hydrostation S has been measured³ using simulated *in situ* ^{14}C incubations, similar to the methods used at BATS today. Mean annual net primary productivity (NPP) from this earlier, negative NAO period was half as high as during the BATS period, 1989–2001. Wintertime maximum NPP was also slightly lower in the earlier years, never exceeding $800 \text{ mg C m}^{-2} \text{ d}^{-1}$, while in the BATS era, NPP exceeded that rate in 7 out of 13 years, and was double it in 1995. This is especially

surprising considering that winter mixed layer depths (MLDs) were 2–4 times deeper in the early years. Whereas the depth of winter mixing has been thought to dictate the magnitude of the winter primary productivity bloom^{4,5}, we found no significant correlation in the BATS data between NPP and winter MLD (as chosen by a change in density from the surface of 0.125 kg m^{-3}), regardless of averaging scheme or lag. Advective changes in the nutrient reservoir may help explain this lack of correlation. The contrast between the two periods studied at Bermuda are consistent with the hypothesis that vigorous STMW production creates a low-nutrient nutristad and exports this signal to the subtropical gyre, thereby reducing NPP downstream of the STMW formation region.

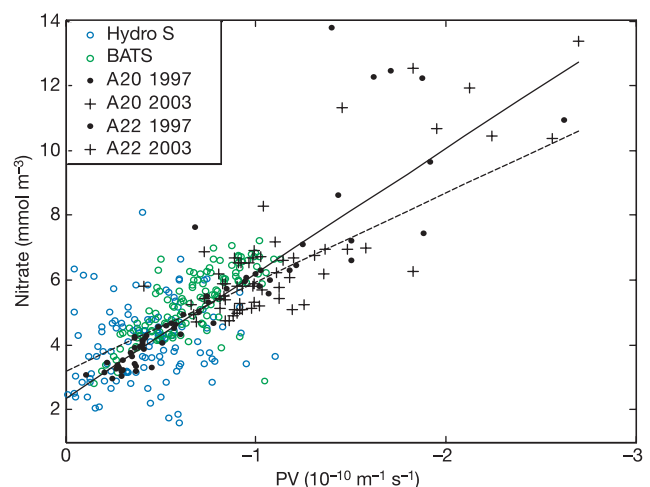


Figure 4 | Nitrate concentration versus PV at the PV minima over the density range $1026.45\text{--}1026.55 \text{ kg m}^{-3}$. Data are shown for the four repeat WOCE sections (black symbols) and the BATS (1988–2001) (green circles) and Hydrostation S (1958–1963) (blue circles) time series data. Gulf Stream data from the WOCE sections are excluded. The solid (dashed) line is the linear regression of the spatial (temporal) data.

Competing mechanisms for nutrient delivery

Satellite observations of sea surface chlorophyll, a measure of phytoplankton standing stock, provide another means of examining the effect of STMW on the biology of the oligotrophic North Atlantic. Large-scale patterns of surface chlorophyll reveal that the subtropical gyre is not uniformly low in chlorophyll. Rather, annual mean chlorophyll constructed from climatology over the length of the SeaWiFS mission (1997–2004) is characterized by a ringed pattern, with the chlorophyll minimum located in the western limb of the central gyre (Fig. 5a). The mechanisms that determine this broad spatial pattern of surface chlorophyll remain poorly understood. To shed light on this issue, we explore qualitatively five mechanisms that possibly affect the spatial pattern of the chlorophyll fields in the subtropical North Atlantic: (1) Ekman downwelling, (2) the strength of the eddy field, (3) the depth of winter mixing, (4) supply by turbulent diffusion, and (5) nutricline displacement by the advection of STMW.

The first mechanism used to explain the location of the chlorophyll minimum is that horizontal convergence of the Ekman transport

creates downwelling such that nutrient-rich, deep water is carried further from the light field¹⁷. This hypothesis suggests that the region of maximum downwelling should approximately coincide with the chlorophyll minimum, but such coincidence is not apparent from an inspection of Fig. 5b. Furthermore, temporal changes in chlorophyll concentration and downwelling velocities contradict the idea that downwelling decreases nutrient availability: the oligotrophic region of the subtropical gyre (chosen as the fraction of the gyre with less than $0.07 \text{ mg chlorophyll m}^{-3}$) shrinks at times of maximum downwelling¹⁷. In agreement with this result, we found that chlorophyll concentration and vertical velocity are negatively correlated across a broad band of the northern subtropical gyre, indicating highest chlorophyll concentrations at times of maximum downwelling. A recent study suggests that this negative correlation is explained by the lateral Ekman flux of inorganic nutrients across the subtropical boundary during strong downwelling¹⁸. While this lateral supply may be important just to the south of the Gulf Stream, possibly reflected by the relatively high chlorophyll at the northern edge of the gyre, it is much reduced towards the centre of the gyre¹⁸.

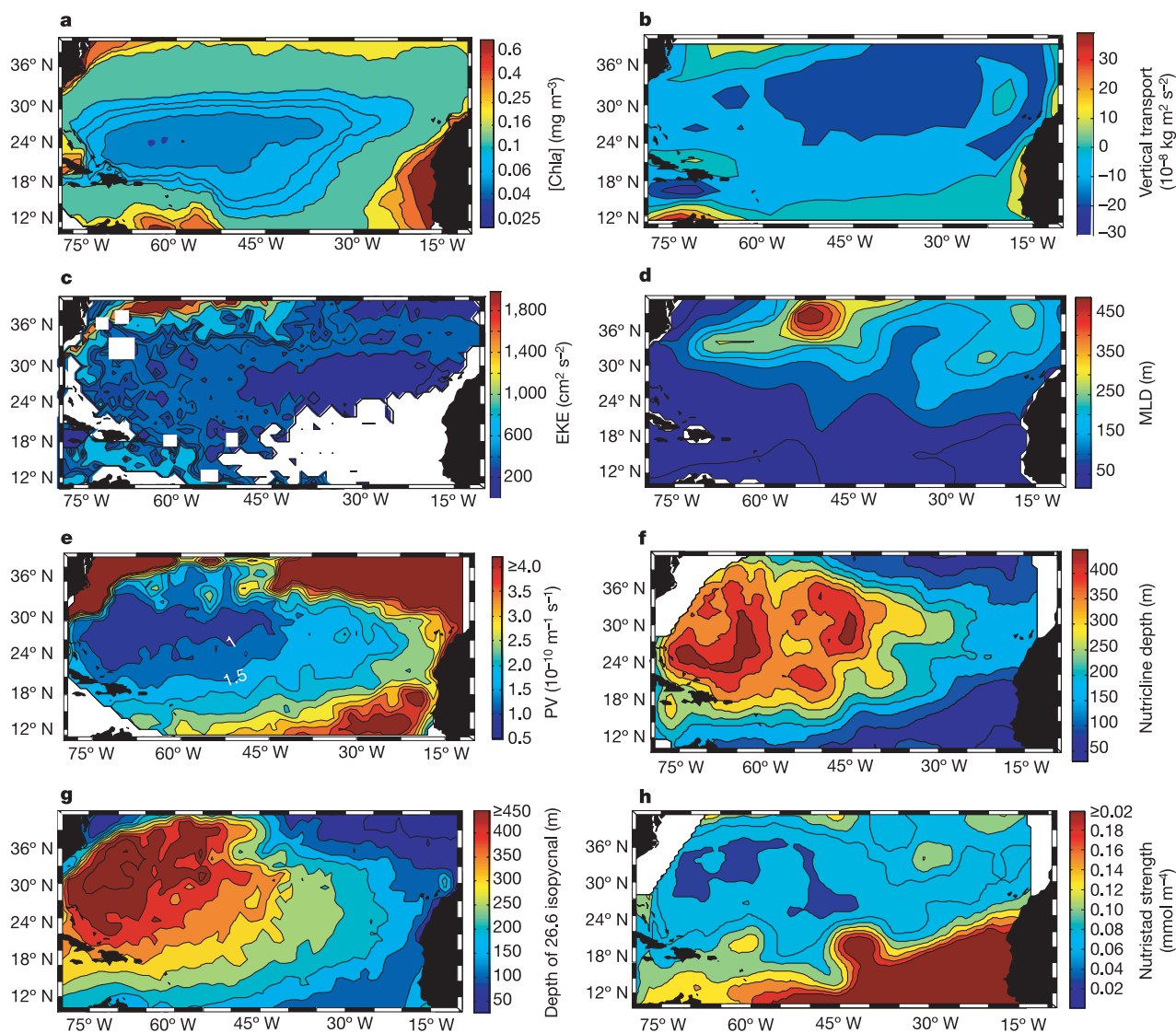


Figure 5 | Properties of the North Atlantic subtropical gyre. **a**, Annual mean SeaWiFS chlorophyll *a* concentration, [Chla], with a log scale for the colour axis; **b**, vertical transport calculated from the annual mean wind stress curl²⁵; **c**, climatological mean eddy kinetic energy²⁰, EKE; **d**, climatological March mixed layer depth²⁶, MLD; **e**, potential vorticity on

the 26.5 isopycnal, PV; **f**, nutricline depth, as defined by the depth of the maximum vertical nitrate gradient; **g**, depth of the 26.6 isopycnal, an approximation for the base of the STMW; and **h**, the strength of the nutrient gradient at the nutricline, showing the wedge of STMW as a depleted nutrientad.

The second mechanism used to explain nutrient availability and spatial patterns of chlorophyll concentration involves mesoscale eddy events¹⁹. A passing eddy can heave a subsurface isopycnal into the euphotic zone and, if the isopycnal is high in nutrients, a surface biological response would result. Hence, it is expected that a strong eddy field, as indicated by the mean eddy kinetic energy (EKE), would correspond with a relatively high phytoplankton standing stock. The map of climatological EKE²⁰ shows a strong meridional gradient that corresponds well with the meridional gradient in satellite-measured chlorophyll (Fig. 5c). However, there is no zonal variation in the EKE field that mimics that of the surface chlorophyll field. To the contrary, EKE tends to be minimized in the eastern part of the subtropical gyre, while satellite chlorophyll is minimized to the west. Furthermore, a recent study²¹ of the relationship between sea level anomalies, used as a measure of thermocline depth changes, and surface chlorophyll found that over much of the subtropical North Atlantic these two are not significantly correlated after the seasonal cycle was removed. The authors suggest that processes other than changes in the thermocline depth are responsible for the observed surface chlorophyll variability. To the extent that the spatial (1°) and temporal (8-day) resolution of the sea level anomaly field allows for the resolution of mesoscale eddies, this study confirms our hypothesis that eddies alone are insufficient in establishing the surface chlorophyll field.

A third mechanism that can bring nutrients to the primary producers is convective mixing^{4,5}. It is clear from the spatial pattern of March MLD that winter mixing is deep enough to penetrate the nutricline within the STMW formation region, and perhaps slightly south of the region (Fig. 5d). Additionally, the spatial pattern of MLD resembles the surface chlorophyll map with regard to the large-scale meridional and zonal gradients (Fig. 5d). However, the closed contours of MLD do not resemble those of surface chlorophyll, and, as noted earlier, no significant temporal correlation between MLD and primary productivity exists in the time series data at Bermuda. We are left concluding that the pattern of convective mixing alone is not sufficient to establish the chlorophyll pattern within the subtropical gyre.

Turbulent diapycnal diffusion is the fourth mechanism examined here that can also supply nutrients to the euphotic zone^{2,22}. A microstructure study conducted south of the Azores², east of the region where STMW resides, found vertical diffusion to be an important source of nitrate to the euphotic zone². However, a study of data from a North Atlantic WOCE section along 24°N showed that upward diffusive nitrate fluxes are much reduced in the western region of the transect where the STMW resides²². Although we lack the data necessary to examine the gyre-scale pattern of turbulent diapycnal diffusion, we infer that the nutrient characteristic of the STMW limits the vertical diffusion of nutrients, as the diffusion must act on a relatively weak gradient.

Having considered a host of vertical processes, we now examine the fifth mechanism that may act as a critical factor in setting the pattern of chlorophyll concentration: the displacement of the nutricline by STMW. The location of the STMW, as indicated by the PV minima on the $\sigma_\theta = 26.5$ (equivalent to $1,026.5 \text{ kg m}^{-3}$) isopycnal (Fig. 5e), coincides with both the region of the deepest nutricline (Fig. 5f) and the chlorophyll minimum (Fig. 5a). Because the nutricline is depressed to accommodate the wedge of STMW, its depth is roughly that of the 26.6 isopycnal (Fig. 5f, g), nominally the base of the STMW. Furthermore, because the wedge of STMW is depleted in nutrients, the strength of the nutrient gradient at the nutricline (Fig. 5h) is weakest within the STMW. The coincidence of this nutrient and the deepened nutricline with the chlorophyll minimum bolsters the hypothesis that the STMW wedge of low-nutrient water in the subtropical gyre acts to limit phytoplankton biomass. As this subsurface water mass establishes the strength of the nutrient reservoir from which all vertical processes draw, spatial patterns in biomass are best explained by a

superposition of the vertical delivery mechanisms and the nutrient reservoir.

Given the role STMW plays in establishing the subsurface nutrient reservoir in the subtropical North Atlantic, understanding the spatial and temporal variability of mode waters throughout the global ocean¹⁰ could shed light on the interannual and decadal changes in global nutrient supply and primary productivity. Thus, the extent to which the mode waters are climatically variable determines the climatic variation of the subsurface nutrient reservoir and, quite possibly, that of a basin's primary productivity.

METHODS

Data sources. Data used for examining STMW spatial variability are from WOCE repeat sections A20 and A22 (Fig. 1), occupied in the summers of 1997 and 2003. Time series data from Hydrostation S (32.10°N, 64.30°W) and BATS (31.92°N, 64.17°W) were both collected approximately biweekly. Hydrostation S nutrient data were collected over the period 1958–63. Primary productivity data from Hydrostation S were taken directly from Menzel and Ryther³. Concern about whether 1960s primary productivity measurements can be compared with modern measurements is warranted, as most primary productivity work before the mid-1980s was systematically low owing to unrecognized trace metal inhibition of phytoplankton productivity²³. However, a comparison of modern primary productivity measurements made with trace-metal clean methods with Menzel and Ryther's primary productivity measurements made using a Teflon and Pyrex water sampler indicated that Menzel and Ryther's method was free from trace metal inhibition²⁴. Thus, we believe that the Menzel and Ryther observations³ can be directly compared with primary productivity observations from BATS.

Satellite chlorophyll data were provided by the SeaWiFS Project, NASA/Goddard Space Flight Center. They are climatological mean values, obtained at 9 km resolution over the length of the mission at the time of writing: September 1997–September 2004. The global ocean wind stress climatology used to calculate Ekman vertical velocities is based on ECMWF (European Centre for Medium-Range Weather Forecasts) analyses²⁵. Spatial fields of EKE were calculated using float trajectories²⁰. March MLD climatology for the North Atlantic was acquired from the Naval Research Laboratory²⁶, with a density criterion based on a change in temperature from the surface of 0.8 °C. Basin-scale hydrographic data was acquired from the National Oceanic Data Center (NODC). Following quality control²⁷, data from 1950–2000 were used to construct climatological PV fields at 1° horizontal resolution.

Preformed nutrients and remineralization rate. Nutrient concentrations in the STMW at the time of subduction were estimated by assuming that the water mass was saturated in oxygen at the time of subduction and that the apparent oxygen utilization (AOU) is caused solely by the remineralization of organic matter according to the Redfield ratio. Thus, the nitrate concentration at the time of subduction, N_s , can be estimated as:

$$N_s = N_m - \text{AOU} \times R_{(\text{NO}_3/\text{O}_2)} \quad (1)$$

where N_m is the measured nitrate concentration, AOU is the difference between the measured oxygen concentration and the saturation oxygen concentration at the observed temperature and salinity, and $R_{(\text{NO}_3/\text{O}_2)}$ is the Redfield ratio of nitrate to oxygen²⁸. Taking N_s at the PV minimum of each WOCE station yields an estimated nitrate (phosphate) concentration at the time of subduction of $1.4 \pm 0.6 \text{ mmol m}^{-3}$ ($0.05 \pm 0.04 \text{ mmol m}^{-3}$). For comparison, Pelegrí and Csanady¹³ show Gulf Stream nitrate concentrations as high as 15 mmol m^{-3} for similar densities at 36°N. CFC ages have been used to calculate remineralization rates by dividing the AOU by the CFC age, and multiplying by the appropriate Redfield ratio. These CFC ages were inferred by comparing CFC-12 concentrations measured along WOCE sections A20 and A22 in 1997 to a time series of atmospheric CFC concentrations²⁹, assuming that the surface water was in equilibrium with the atmosphere before subduction³⁰. The correlation between CFC age and PV for both of these sections was calculated with data at the PV minima from 20–38°N, where the STMW resides.

Scale analysis. To compare advective, diffusive and remineralization timescales, we consider the conservation equation for nitrate:

$$\frac{\partial N}{\partial t} + \mathbf{u} \cdot \nabla N = k_H \nabla_H^2 N + k_V \frac{\partial^2 N}{\partial z^2} + R \quad (2)$$

where N is the nitrate concentration, $\mathbf{u}$ is the velocity of a fluid parcel, k_H and k_V are the horizontal and vertical diffusivities, and R is the source of nutrients due to remineralization. The Peclet number, UD^2/Lk_V , is the ratio of the vertical diffusive timescale to the horizontal advective timescale. (Here D is the depth scale, L is the length scale, and U is the horizontal velocity scale.) This ratio is of the order of 10^3 using a k_V value of $10^{-5} \text{ m}^2 \text{ s}^{-1}$, as calculated in tracer release

experiments³¹; a horizontal speed of 10 cm s^{-1} ; and appropriate values of L and D for the depleted nutricline (2,000 km and 500 m, respectively). This high Peclet number reflects the dominance of horizontal advection over diffusion in setting the nitrate concentration. A similar analysis shows that along-isopycnal or horizontal diffusion is also relatively weak. The ratio of the remineralization timescale to the advective timescale, UN/LR , is calculated with the same characteristic horizontal velocity and length scale as above, a nitrate concentration (N) of 1 mmol m^{-3} , and a remineralization rate (R) of $0.53\text{--}1.5 \text{ mmol m}^{-3} \text{ yr}^{-1}$. We calculated this lower bound on the remineralization rate, as explained above. The upper bound was inferred from a previous study of oxygen utilization rates in the North Atlantic³². The resulting remineralization to advective timescale ratio ranges from 1 to 5.

Received 24 January; accepted 17 June 2005.

- Mann, K. H. & Lazier, J. R. N. *Dynamics of Marine Ecosystems* (Blackwell Science, Cambridge, Massachusetts, 1996).
- Lewis, M. R., Harrison, W. G., Oakey, N. S., Hebert, D. & Platt, T. Vertical nitrate fluxes in the oligotrophic ocean. *Science* **234**, 870–873 (1986).
- Menzel, D. W. & Ryther, J. H. Annual variations in primary productivity of the Sargasso Sea off Bermuda. *Deep-Sea Res.* **17**, 282–288 (1961).
- Gruber, N., Keeling, C. D. & Bates, N. R. Interannual variability in the North Atlantic Ocean carbon sink. *Science* **298**, 2374–2378 (2002).
- Lewis, M. R., Kuring, N. & Yentsch, C. Global patterns of ocean transparency: Implications for the new production of the open ocean. *J. Geophys. Res.* **93**, 6847–6856 (1988).
- McGillicuddy, D. J. & Robinson, A. R. Eddy-induced nutrient supply and new production in the Sargasso Sea. *Deep-Sea Res.* **44**, 1427–1450 (1997).
- Oschlies, A. & Garçon, V. Eddy-induced enhancement of primary production in a model of the North Atlantic Ocean. *Nature* **394**, 266–269 (1998).
- Sarmiento, J. L., Gruber, N., Brzezinski, M. A. & Dunne, J. P. High-latitude controls of thermocline nutrients and low latitude biological productivity. *Nature* **427**, 56–60 (2004).
- Talley, L. D. & Raymer, M. E. Eighteen Degree Water variability. *J. Mar. Res.* **40** (suppl.), 757–775 (1982).
- McCartney, M. S. The subtropical recirculation of Mode Waters. *J. Mar. Res.* **40**, 427–464 (1982).
- Worthington, L. V. *On the North Atlantic Circulation* (Johns Hopkins University Press, Baltimore, Maryland, 1976).
- Siegel, D. A., Doney, S. C. & Yoder, J. A. The North Atlantic spring phytoplankton bloom and Sverdrup's critical depth hypothesis. *Science* **296**, 730–733 (2002).
- Pelegri, J. L. & Csanady, G. T. Nutrient transport and mixing in the Gulf Stream. *J. Geophys. Res.* **96**, 2577–2583 (1991).
- Joyce, T., Deser, C. & Spall, M. A. The relation between decadal variability of subtropical mode water and the North Atlantic Oscillation. *J. Clim.* **13**, 2550–2569 (2000).
- Talley, L. D. North Atlantic circulation and variability, reviewed for the CNLS conference. *Physica D* **98**, 625–646 (1996).
- Dickson, R., Lazier, J., Meincke, J., Rhines, P. & Swift, J. Long-term coordinated changes in the convective activity of the North Atlantic. *Prog. Oceanogr.* **38**, 241–295 (1996).
- McClain, C. R., Signorini, S. R. & Christian, J. R. Subtropical gyre variability observed by ocean-colour satellites. *Deep-sea Res.* **51**, 281–301 (2004).
- Williams, R. G. & Follows, M. J. The Ekman transfer of nutrients and maintenance of new production over the North Atlantic. *Deep-Sea Res.* **45**, 461–489 (1998).
- McGillicuddy, D. J. *et al.* Influence of mesoscale eddies on new production in the Sargasso Sea. *Nature* **419**, 263–266 (1998).
- Fratantoni, D. M. North Atlantic surface circulation during the 1990's observed with satellite-tracked drifters. *J. Geophys. Res.* **106**, 22067–22093 (2001).
- Wilson, C. & Adamec, D. A global view of bio-physical coupling from SeaWiFS and TOPEX satellite data, 1997–2001. *Geophys. Res. Lett.* **29**, doi:10.1029/2001GL014063 (2002).
- Bahamón, N., Velasquez, Z. & Cruzado, A. Chlorophyll *a* and nitrogen flux in the tropical North Atlantic Ocean. *Deep-Sea Res.* **50**, 1189–1203 (2003).
- Fitzwater, S., Knauer, G. A. & Martin, J. H. Metal contamination and its effect on primary production measurements. *Limnol. Oceanogr.* **27**, 544–551 (1982).
- Barber, R. T. *et al.* Primary productivity and its regulation in the Arabian Sea during 1995. *Deep-Sea Res.* **48**, 1127–1172 (2001).
- Trenberth, K. E., Olsen, J. G. & Large, W. G. *A Global Ocean Wind Stress Climatology Based on ECMWF Analyses* (Tech. Note NCAR/TN-338 + STR, National Center for Atmospheric Research, Boulder, Colorado, 1989).
- Kara, A. B., Rochford, P. A. & Hurlburt, H. E. Mixed layer depth variability over the global ocean. *J. Geophys. Res.* **108**, doi:10.1029/2000JC000736 (2003).
- Lozier, M. S., Owens, W. B. & Curry, R. G. The climatology of the North Atlantic. *Prog. Oceanogr.* **36**, 1–44 (1995).
- Takahashi, T., Broecker, W. S. & Langer, S. Redfield ratio based on chemical data from isopycnal surfaces. *J. Geophys. Res.* **90**, 6907–6924 (1985).
- Walker, S. J., Weiss, R. F. & Salameh, P. K. Reconstructed histories of the annual mean atmospheric mole fractions for the halocarbons CFC-11, CFC-12, CFC-113, and carbon tetrachloride. *J. Geophys. Res.* **105**, 14285–14296 (2000).
- Warner, M. J. & Weiss, R. F. Solubilities of chlorofluorocarbons 11 and 12 in water and seawater. *Deep-Sea Res.* **32**, 1485–1497 (1985).
- Ledwell, J. R., Watson, A. J. & Law, C. S. Evidence for slow mixing across the pycnocline from an open-ocean tracer-release experiment. *Nature* **364**, 701–703 (1993).
- Jenkins, W. J. Oxygen utilization rates in North Atlantic subtropical gyre and primary production in oligotrophic systems. *Nature* **300**, 246–248 (1982).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Lethaby for the Hydrostation S data, D. Fratantoni for the EKE data, and D. LaBel and W. Smethie for the CFC age data. This Article also benefited from discussions with M. Follows and W. Jenkins. This work was supported by an NSF Graduate Research Fellowship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.B.P. (jbp3@duke.edu).

GIBBERELLIN INSENSITIVE DWARF1

encodes a soluble receptor for gibberellin

Miyako Ueguchi-Tanaka^{1*}, Motoyuki Ashikari^{1*}, Masatoshi Nakajima^{2*}, Hironori Itoh¹, Etsuko Katoh³, Masatomo Kobayashi⁴, Teh-yuan Chow^{5†}, Yue-ie C. Hsing⁵, Hidemi Kitano¹, Isomaro Yamaguchi^{2,6} & Makoto Matsuoka¹

Gibberellins (GAs) are phytohormones that are essential for many developmental processes in plants. It has been postulated that plants have both membrane-bound and soluble GA receptors; however, no GA receptors have yet been identified. Here we report the isolation and characterization of a new GA-insensitive dwarf mutant of rice, *gid1*. The *GID1* gene encodes an unknown protein with similarity to the hormone-sensitive lipases, and we observed preferential localization of a GID1-green fluorescent protein (GFP) signal in nuclei. Recombinant glutathione S-transferase (GST)-GID1 had a high affinity only for biologically active GAs, whereas mutated GST-GID1 corresponding to three *gid1* alleles had no GA-binding affinity. The dissociation constant for GA₄ was estimated to be around 10⁻⁷ M, enough to account for the GA dependency of shoot elongation. Moreover, GID1 bound to SLR1, a rice DELLA protein, in a GA-dependent manner in yeast cells. GID1 overexpression resulted in a GA-hypersensitive phenotype. Together, our results indicate that GID1 is a soluble receptor mediating GA signalling in rice.

Gibberellins (GAs) are a large family of tetracyclic, diterpenoid plant hormones that induce a wide range of plant growth responses including seed germination, stem elongation, leaf expansion, pollen maturation and induction of flowering¹. Although the biosynthesis of GA has been well characterized^{2,3}, little is known about how plants perceive GA and how the GA signal is transmitted to cause GA-regulated plant growth.

A rice mutant, *slender rice1* (*slr1*), shows a constitutive GA response phenotype^{4,5}. The *SLR1* gene encodes a putative transcriptional regulator orthologous to *Arabidopsis* GAI⁶ and RGA⁷, wheat Rht, maize D8 and barley SLN⁸. These proteins are referred to as the DELLA subfamily of the GRAS regulatory protein family because they share a conserved sequence called DELLA. DELLA proteins have been suspected to function as suppressors of GA signalling because their degradation triggers various GA responses *in planta*⁹. Recently, we isolated and characterized a rice GA-insensitive dwarf mutant, *gid2* (ref. 10). The *GID2* gene encodes a putative F-box subunit of an SCF E3 ubiquitin ligase that interacts with a rice Skp1 homologue in the yeast two-hybrid assay¹¹. Moreover, high levels of SLR1 accumulation were observed in the *gid2* mutants. On the basis of these observations, we proposed that GA treatment induces the degradation of SLR1 through the SCF^{GID2}-proteasome pathway.

However, molecular mechanisms of GA perception have yet to be clarified. The biochemical properties of GA, which is a hydrophobic carboxylic acid, indicate that it is soluble in the intercellular and intracellular compartment of plant cells as a carboxylate anion, and that it may cross the plasma membrane by passive diffusion as a protonated acid¹². Therefore, it has been postulated that plants have both membrane-bound and soluble GA receptors. Although there are some reports of detection of GA-binding proteins by biochemical approaches^{13,14}, these proteins have not been isolated, nor is there

corroborating genetic evidence that these binding proteins act as GA receptors.

In order to investigate the GA signalling mechanism, we screened rice *gid* mutants and identified several *gid* mutations at different loci.

Rice *gid1* mutant has a GA-insensitive phenotype

One of the mutants, *gid1-1*, had a severe dwarf phenotype with wide, dark-green leaf blades (Fig. 1a), typical of rice GA-related mutants^{3,10,15}. The mutant was inherited in a recessive manner and did not develop fertile flowers; thus it had to be maintained as a heterozygote. So far, we have isolated four different alleles. Three of them, *gid1-1*, *gid1-3* and *gid1-4*, show similarly severe dwarfism, whereas the remaining one, *gid1-2*, has a slightly milder phenotype than the others (data not shown). *gid1-1* plants do not exhibit any of the GA-responsive phenotypes we examined, including elongation of the second leaf sheath (Fig. 1b) and induction of α -amylase activity in seeds (Fig. 1c). Negative feedback¹⁶ in the expression of the GA biosynthetic gene *SD1* (also known as *OsGA20ox2*, ref. 17) by GA₃ was observed in wild type but not in *gid1-1* plants (Fig. 1d). We also measured the endogenous levels of GAs and found that *gid1-1* and *gid1-2* accumulate at about 120 (Fig. 1e) and 95 times (data not shown) the level of GA₁ found in wild-type plants, respectively. These results demonstrated that *gid1* is a GA-insensitive mutant.

SLR1 is epistatic to GID1 and is not degraded in the *gid1* mutant

A *gid1-1/slr1-1* double mutant exhibited the *slr1-1* phenotype (Fig. 2a), indicating that GID1 and SLR1 function in the same GA signalling pathway and that SLR1 is epistatic to GID1. GA-dependent degradation of SLR1 is essential for GA action, and if degradation is inhibited, plants show the GA-insensitive phenotype¹⁰. Immunoblot analysis of the SLR1 protein showed that GA₃ treatment induced

¹Bioscience and Biotechnology Center, Nagoya University, Nagoya 464-8601, Japan. ²Department of Applied Biological Chemistry, The University of Tokyo, Tokyo 113-8657, Japan. ³Department of Biochemistry, National Institute of Agrobiological Sciences, Tsukuba 305-8602, Japan. ⁴BioResources Center, Riken, Tsukuba 305-0074, Japan. ⁵Institute of Botany Academia Sinica, Taipei 11529, Taiwan. ⁶Biotechnology Research Center, The University of Tokyo, Tokyo 113-8657, Japan. [†]Present address: Institute of Biotechnology, Central Taiwan University of Science and Technology, Taichung 406, Taiwan.

*These authors contributed equally to this work.

complete degradation of SLR1 within 30 min in the wild type (Fig. 2b, top panel). In *gid1-1*, SDS–polyacrylamide gel electrophoresis gave two bands corresponding to the non-phosphorylated (lower band) and phosphorylated (upper band) forms of SLR1 (refs 10, 18; Fig. 2b, top panel). The intensity of these bands was stronger in *gid1-1* plants than in wild-type plants. Furthermore, GA treatment did not diminish the amount of SLR1 in *gid1-1* plants whereas it did so in the wild type (Fig. 2b, top panel). Enhanced SLR1 stability in *gid1-1* mutants was also confirmed in transgenic plants producing SLR1 promoter–SLR1–GFP. The GFP signal was observed in nuclei of *gid1-1* cells after GA₃ treatment, but not in wild-type nuclei (Fig. 2b, bottom panels), demonstrating that GID1 is essential for SLR1 degradation.

gid1 resembles *cps* but not *gid2* in SLR1 accumulation

The failure of *gid1-1* mutants to degrade SLR1 (Fig. 2b) suggests that GID1 is either directly involved in the degradation of SLR1, as is GID2 (ref. 10), or that it acts earlier in the GA signalling cascade. In order to investigate the position of GID1 in the GA signalling cascade, we compared the phenotype of *gid1-3* to that of GA-related

mutants such as *gid2-2*, which is defective in SLR1 degradation, and *cps*, which is defective in the GA synthesis enzyme copalyl diphosphate synthase. Although we found dwarfism to be less severe in *gid2-2* than in *cps* and *gid1-3* (Fig. 2c, top panel), the amount of SLR1 accumulated in *gid2-2* was much higher than in *cps* and *gid1-3* (Fig. 2c, middle panel). This indicates that the SLR1-dependent suppression of GA action is weaker in *gid2-2* than in *cps* and *gid1-3* mutants. Because the *cps* mutant is unable to produce active GAs, GA signalling cannot be initiated endogenously in this mutant. In contrast, the GA signal will be able to reach SLR1 in *gid2-2*, but no degradation of SLR1 will occur in this mutant. Therefore we speculate that the SLR1-dependent suppression of GA response may be regulated by GA itself, and that SLR1 may be less effective in *gid2-2* compared to *cps*, where there is no endogenous GA. Because the phenotype of *gid1-3* was similar to *cps* but not to *gid2-2* with respect to dwarfism and SLR1 accumulation, the GA signal may not reach SLR1 in *gid1-3* mutants. Thus, GID1 may have a function in the perception of GA rather than in SLR1 degradation.

GID1 encodes an unknown protein with similarity to HSLs

To understand the molecular function of GID1, we isolated *GID1* by positional cloning (Supplementary Fig. S1). The *GID1* gene contains one intron and two exons, and encodes a 354-amino-acid polypeptide (Fig. 3a, b), which was confirmed by sequencing full-length

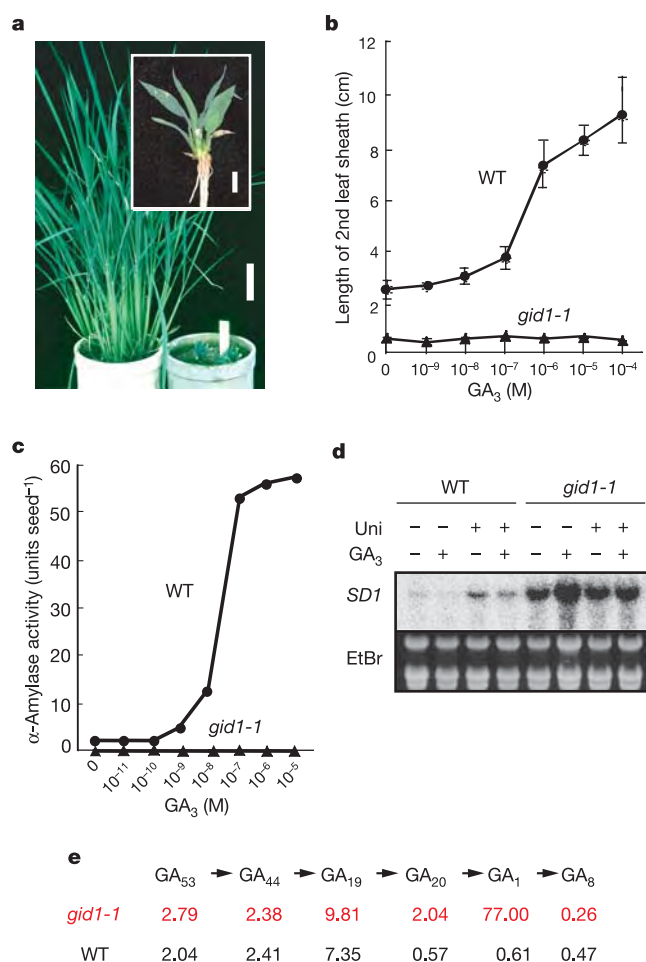


Figure 1 | GA-insensitive phenotype of *gid1-1*. **a**, Gross morphology of wild-type (left) and *gid1-1* (right) plants. Scale bar, 10 cm. Inset: higher magnification of *gid1-1*. Scale bar, 1 cm. **b**, GA₃-induced elongation of the second leaf sheath (mean ± s.d.; n = 10). **c**, GA₃ induction of α-amylase activity in embryoless half seeds. **d**, RNA gel blot analysis of *SD1* (also known as *OsGA20ox2*). Total RNA was extracted from seedlings grown with (+) or without (–) 10⁻⁶ M uniconazol (Uni), a GA biosynthesis inhibitor, for 2 weeks and then treated with (+) or without (–) 10⁻⁵ M GA₃. EtBr shows rRNA bands as a loading control. **e**, GA levels (ng per gram fresh weight) in the wild type (WT) and *gid1-1* mutant. Representative results from one of three independent experiments are shown.

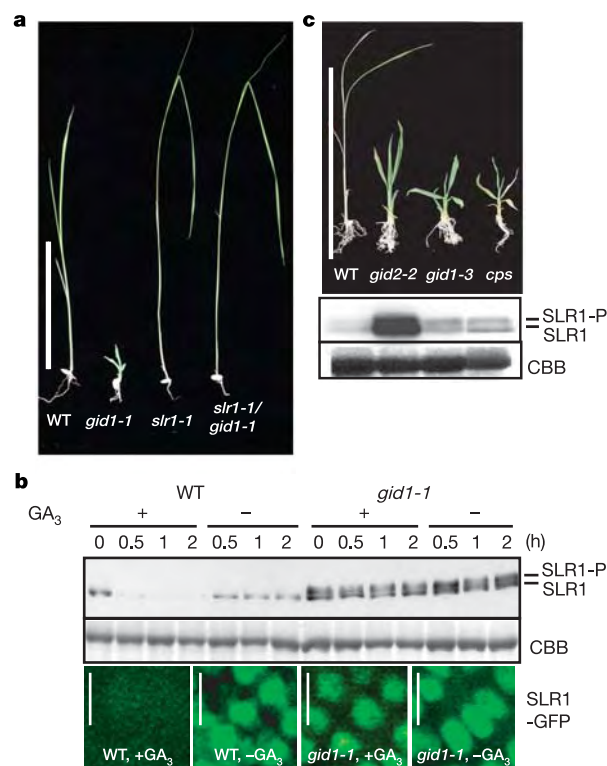


Figure 2 | The *slr1-1/gid1-1* double mutant has the *slr1-1* phenotype.

a, Epistatic analysis of *gid1-1* and *slr1-1* mutants. Scale bar, 10 cm. **b**, GA-mediated SLR1 degradation. Top panel: western blot analysis of SLR1. Two-week-old seedlings were grown with 10⁻⁶ M uniconazol and then treated with 10⁻⁴ M GA₃ for the period indicated. Ten micrograms of total protein was applied in each lane. SLR1-P, phosphorylated SLR1; SLR1, non-phosphorylated SLR1. Middle panel: CBB control. Bottom panel: GFP fluorescence in the wild type and *gid1-1* mutant carrying SLR1 promoter–SLR1–GFP. Plants were grown with 10⁻⁶ M uniconazol and then treated with (+) or without (–) 10⁻⁴ M GA₃ for 12 h. Scale bars, 5 μm. **c**, Comparison between wild type, *gid1-3*, *gid2-2* and *cps*. Top panel: gross morphology. Scale bar, 10 cm. Middle panel: western blot analysis of SLR1. Ten micrograms of total protein was applied in each lane. Bottom panel: CBB control.

complementary DNA. The four *gid1* alleles had single-nucleotide substitutions in the coding region (*gid1-1* and *gid1-2*) or internal deletions between intron 1 and exon 2 (*gid1-4*) or in exon 2 (*gid1-3*) (Fig. 3a, b). Introduction of the *GID1*-containing 6.7-kilobase (kb) *Pst*I genomic fragment into the *gid1-1* mutant restored the normal phenotype (data not shown). A database search revealed that there is no gene homologous to *GID1* in rice, whereas there are three homologues in *Arabidopsis*, which are annotated as unknown proteins. An NCBI Conserved Domain Search indicated that *GID1* shares homology with the consensus sequence of the hormone-sensitive lipase (HSL) family¹⁹ (Fig. 3c), including the conserved HSL motifs HGG and GXSGX^{20,21} (filled circles in Fig. 3c). The importance of the GXSGX motif is highlighted by the severe

phenotype of *gid1-1*, a single amino acid exchange mutant in which the first G of this motif is replaced by D (Fig. 3b). Three conserved amino acids, S, D and H, form the catalytic triad in the HSL family²² (filled squares in Fig. 3c). Two of them, S and D, were also conserved in *GID1*, whereas the third, H, was replaced by V. As this amino acid, H, is essential for HSL catalytic function²², *GID1* should be expected to lack the enzyme activity. In fact, recombinant *GID1* did not hydrolyse *p*-nitrophenyl acetate (data not shown), which is an artificial substrate for HSL proteins. Transgenic *GID1*-GFP protein expressed under the control of the rice actin promoter (pAct1) was primarily localized in nuclei, with a fainter cytosolic signal (Fig. 3d). This cellular localization did not change with uniconazol or GA₃ treatment.

GID1 is a soluble GA receptor

To examine the postulated involvement of *GID1* in the perception of GA, we studied the interaction between recombinant GST-*GID1* and radioactive GA using a non-equilibrium gel-permeation technique¹⁴. GST-*GID1* bound to [1,2,16,17-³H₄]16,17-dihydro-GA₄ (³H₄-16,17-dihydro-GA₄), and most of the binding was replaceable with excess unlabelled GA₄, indicating that the competition was

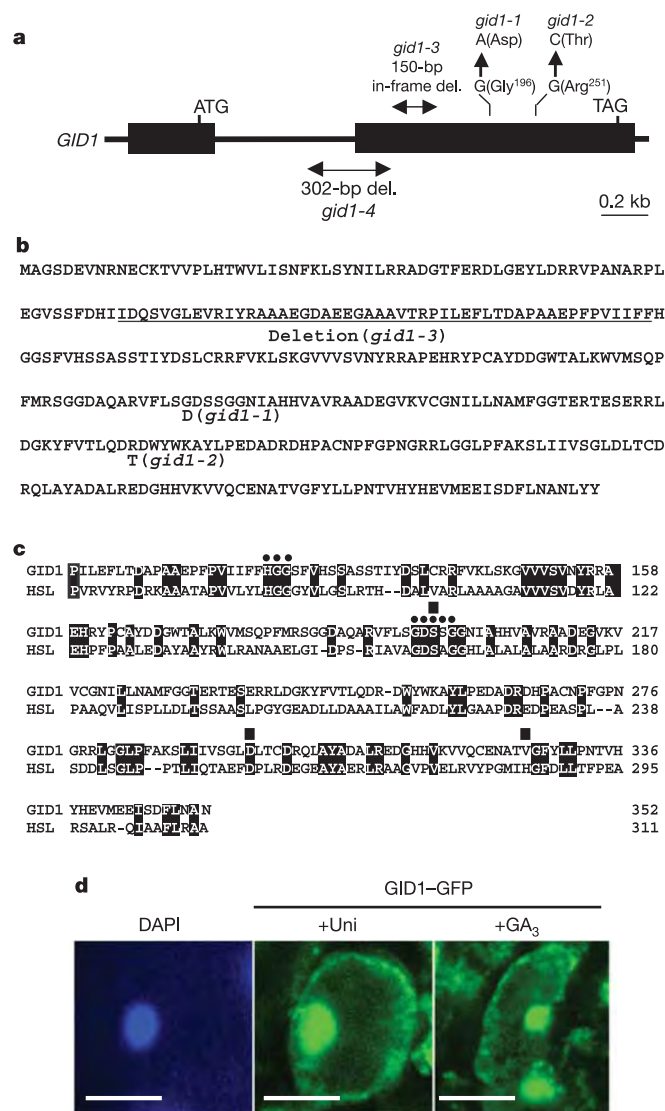


Figure 3 | Structure of *GID1*. **a**, Structure of the *GID1* gene and its mutation sites in the four *gid1* alleles. The *GID1* gene consists of two exons (thick lines) and one intron (thin line). Nucleotide deletions and substitutions in the four *gid1* alleles are indicated. **b**, The deduced amino acid sequence of *GID1*. The *gid1-1*, *gid1-2* and *gid1-3* mutations are also indicated. **c**, Comparison of amino acid sequences between *GID1* and the HSL consensus sequence. Filled circles and squares represent conserved regions and the catalytic triad in the HSL family, respectively. Numbers indicate the position from the start codon. **d**, GFP fluorescence in leaf sections of transgenic rice carrying actin1 promoter-*GID1*-GFP. Plants were treated with 10^{-6} M uniconazol (+Uni) or 10^{-5} M GA₃ (+GA₃) for 1 week. The left panel represents DAPI staining of the central image. Scale bars, 5 μ m.

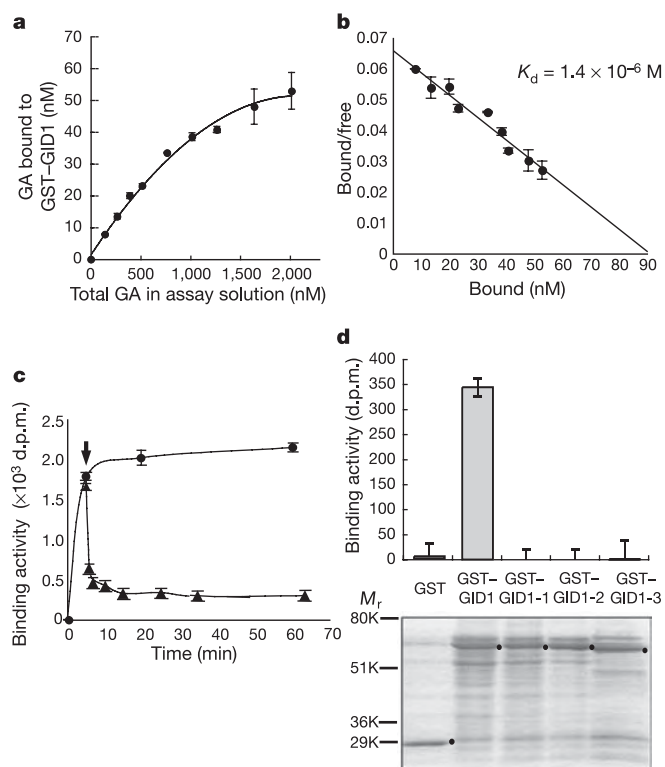


Figure 4 | GA-binding properties of *GID1*. **a**, GA-binding saturability of GST-*GID1*. GST-*GID1* was incubated with 6 pmol ³H₄-16,17-dihydro-GA₄ and increasing concentrations of unlabelled 16,17-dihydro-GA₄ (mean $\pm$ s.d.; $n = 3$). Total binding of 16,17-dihydro-GA₄ (labelled plus unlabelled) was calculated from labelled ligand binding. **b**, Scatchard plot of binding data in **a**. K_d values were calculated from three independent experiments ($R^2 = 0.96$). Data are mean $\pm$ s.d., $n = 3$. **c**, Association/dissociation rates of ³H₄-16,17-dihydro-GA₄ and GST-*GID1*. Total binding of ³H₄-16,17-dihydro-GA₄ reached one-half of the maximum within 5 min (filled circles). Addition of the unlabelled GA₄ (0.125 mM, arrow) reduced ³H₄-16,17-dihydro-GA₄ binding to less than 10% within 5 min (filled triangles). d.p.m., disintegrations per minute. Data are mean $\pm$ s.d.; $n = 3$. **d**, Top panel: the three mutated GST-*GID1* proteins (GST-*GID1*-1, GST-*GID1*-2 and GST-*GID1*-3) did not interact with GA₄. GST, GST tag alone. Data are mean $\pm$ s.d.; $n = 3$. Bottom panel: CBB control. Dots indicate the GST-*GID1* proteins or GST tag alone on SDS-PAGE. Approximately equal amounts of protein (about 3.2 μ g) were used for the assay.

GA specific (Supplementary Fig. S2a, b). Neither heat-denatured GST–GID1 nor GST–GID2, an F-box protein involved in GA-signalling¹⁰, had any specific binding activity (Supplementary Fig. S2a, b). The GA-binding activity of native GID1 protein, with the GST tag cleaved off and subsequently purified, was also detected and it was slightly higher than that of GST–GID1 (Supplementary Fig. S2a, c). The higher binding activity of native GID1 may be due to a greater specific amount of purified native GID1 in the assay solution.

We performed a kinetic analysis of GST–GID1 binding to GA by determining binding saturability with various concentrations of 16,17-dihydro-GA₄ (Fig. 4a). Scatchard plot analysis revealed that the dissociation constant (K_d) for 16,17-dihydro-GA₄ was 1.4×10^{-6} M (Fig. 4b). We also examined the ligand specificity of GST–GID1 by competition between ³H₄-16,17-dihydro-GA₄ and ten GAs with differing biological activity (Supplementary Fig. S3). Table 1 lists the concentration of each GA required for 50% inhibition (IC₅₀) of ³H₄-16,17-dihydro-GA₄ binding to GST–GID1. GST–GID1 showed high affinity for biologically active GAs such as GA₄, 16,17-dihydro-GA₄, GA₁ and GA₃, whereas it had low affinity, or none at all, for biologically inactive GAs. These IC₅₀ values were generally consistent with the physiological activity of the different GAs²³, with the exception that the IC₅₀ value of GA₄ was smaller than that of GA₃, although the physiological activity of GA₄ is lower than that of GA₃. Notably, GA₃ has a double bond at the 2'-carbon (Supplementary Fig. S3) that prevents GA₃ inactivation through GA 2-oxidase²⁴. In contrast to GA₃, GA₄ should be inactivated by GA 2-oxidase *in planta*, which explains the apparently low physiological activity of GA₄.

Because ³H₄-16,17-dihydro-GA₄ was the only tritiated ligand available, we were unable to establish K_d values for other GAs. Assuming that K_d values are similar to IC₅₀, we estimate K_d values for GA₃ and GA₄ to be about 4×10^{-6} M and 2×10^{-7} M, respectively. The estimated K_d value for GA₃ seems to be consistent or slightly lower than the 50% response point of the dose–response curve of GA-induced leaf elongation (Fig. 1b, see below).

The half-time for both association and dissociation between GST–GID1 and 16,17-dihydro-GA₄ was within 5 min (Fig. 4c), indicating that these reactions occur very quickly. Rapid GA-binding kinetics may be critical for soluble receptors because of the sensitivity of the system to subtle alterations in intracellular GA concentrations, which in turn have profound and compounding effects on gene regulation. Such rapid receptor–ligand kinetics have also been reported for mammalian soluble receptors²⁵. The mutated GST–GID1 proteins corresponding to the three alleles of the *gid1* mutant (*gid1-1*, *gid1-2* and *gid1-3*) had no GA-binding activity (Fig. 4d). Thus, the single amino acid substitutions in GST–GID1-1 and GST–GID1-2, and the deletion in GST–GID1-3, cause a loss of GA-binding ability, resulting in GA insensitivity.

Table 1 | Competition for ³H₄-16,17-dihydro-GA₄ binding to GID1 by GAs

GAs	IC ₅₀	Relative percentage
Biologically active GAs*		
GA ₄	2×10^{-7} M	100
H ₂ -GA ₄	1×10^{-6} M	20
GA ₁	4×10^{-6} M	5
GA ₃	4×10^{-6} M	5
Weakly biologically active GAs*		
GA ₃₅	1×10^{-5} M	2
GA ₃₇	2×10^{-5} M	1
Biologically inactive GAs*		
GA ₄ -Me	3×10^{-5} M	0.6
GA ₉	2×10^{-4} M	0.1
GA ₅₁	$>2 \times 10^{-4}$ M	<0.1
3- <i>epi</i> -GA ₄	$>2 \times 10^{-4}$ M	<0.1

H₂-GA₄, 16,17-dihydro-GA₄; GA₄-Me, GA₄ methyl ester.

* Biological activities for various GAs are taken from ref. 23.

If GID1 is a GA receptor, perception of GA by GID1 should be transduced to SLR1, a downstream component of the GA signalling pathway. We tested this prediction using a yeast two-hybrid assay. GID1 interacted with SLR1 in yeast cells in the presence of GA₃, but not in GA₃-free medium (Fig. 5a). This indicates that GID1 interacts directly with SLR1 in a GA-dependent manner and probably transduces the GA signal to SLR1, resulting in SLR1 degradation. We also generated transgenic rice plants overproducing GID1 controlled by pAct1, to see whether such plants show a GA-hypersensitive phenotype. The plants were tall with long, light-green leaves, fewer tillers and poor fertilities compared with the control plants, all of which is consistent with a GA overdose phenotype (Fig. 5b). Using the growth of the second leaf sheath as an index, the sensitivity of these lines to GA application was found to be about ten times higher than that of control plants at the 50% response point of the dose–response curve (Fig. 5c).

Discussion

We have described the cloning and characterization of a new GA signal-related gene, *GID1*, from rice. We conclude that GID1 is a soluble GA receptor from the following evidence: (1) loss-of-function mutations in *GID1* produce a severe dwarf phenotype with loss of GA responsiveness; (2) GST–GID1 interacts with biologically active GAs but not with inactive GAs with reasonable dissociation constants; (3) mutated GST–GID1 proteins corresponding to three *gid1* alleles lack GA binding; (4) GID1 interacts with a rice DELLA protein, SLR1, in a GA-dependent manner in transformed yeast cells; and (5) overexpressors of GID1 show a GA-hypersensitive phenotype.

Although GA-binding proteins have been described before, there has been no genetic or biochemical evidence to link them with signal transduction, making this the first identification, to our knowledge, of a soluble GA receptor. However, our findings do not rule out the possibility of an alternative membrane-bound GA receptor. The GA-dependent induction of α -amylase in the aleurone layer shows a

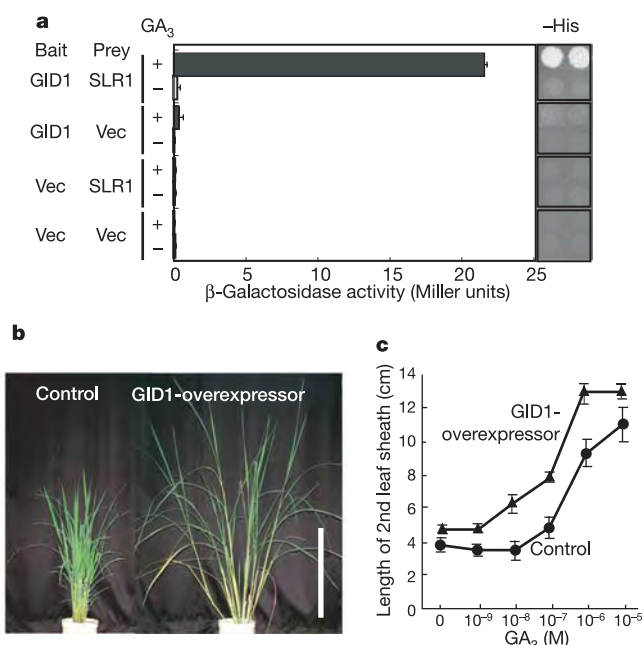


Figure 5 | Interaction between GID1 and SLR1, and the phenotype of a GID1 overexpressor. a, GID1 binds to SLR1 in a GA-dependent manner. Left: β -galactosidase activity detected in a liquid assay with Y187 transformants (mean $\pm$ s.d.; $n = 3$). Right: growth of HA109 transformants on a -His plate. **b**, Gross morphology of GID1-overexpressor and control plants. Scale bar, 50 cm. **c**, Elongation of the sheath of the second leaf with exogenous GA₃ (mean $\pm$ s.d.; $n = 5$).

higher sensitivity to GA₃ than GA-induced leaf sheath elongation (Fig. 1c), suggesting that alternative receptors or other factors cooperate with GID1 in aleurone cells^{12,26}.

GID1 shares conserved motifs with the HSL family of proteins. At present, we can only speculate about the functional implications and evolutionary significance of these similarities. The conserved motifs of the HSL family are crucial for the substrate–enzyme interactions of these proteins. It seems possible that these motifs also mediate the interaction of GID1 with GA. Structural analyses of GID1 will help to reveal its molecular mechanism as a GA receptor.

On the basis of previous and current observations, we propose the following model of GA signalling. A soluble GA receptor, GID1, is present in the nucleus. When GID1 binds a biologically active GA molecule, it attains the ability to interact with SLR1. As a consequence of this interaction, SLR1 becomes degradable through the SCF^{GID2} proteasome pathway. At present, the detailed molecular mechanisms underlying this proposed scheme remain obscure. In particular, it is unclear whether GA–GID1 induces stable conformational changes in SLR1 that render the latter accessible to the SCF^{GID2} complex, or whether the GA–GID1–SLR1 complex as a whole is targeted by SCF^{GID2}. The existence of such a relatively stable GA–GID1–SLR1 complex might be indicated by the fact that the *in vitro* binding affinity of GST–GID1 to GA₃ seems to be slightly too low to account for the GA dose-response curve of leaf elongation.

Recently, the F-box protein TIR1 was identified as an auxin receptor^{27,28}. TIR1 directly interacts with auxin to promote its conformational change that favours Aux/IAA binding. GA and auxin signalling mechanisms seem to be similar at least with regard to the perception of the ligands, as both GA and auxin are bound by nuclear soluble receptors and because ligand perception leads to degradation of negative regulators (SLR1 for GA and Aux/IAA proteins for auxin) through the SCF-mediated 26S proteasome system. Surprisingly, a direct, ligand-dependent interaction of the receptors with negative regulators forms a central step in both signalling cascades, although the receptors differ structurally as they are members of the HSL-like protein family and the F-box protein family, respectively. This type of signal transduction system is unusual in animal cells, and it will be interesting to see whether plant cells also use it in other signal transduction pathways.

METHODS

Plant materials and growth conditions. A Japonica-type rice cultivar (*Oryza sativa* L. cv. Taichung 65) and its chemically or irradiation-induced mutant lines with the alleles *gid1-1* to *gid1-4* and *slr1-1* (ref. 4) were used in this study. To generate *gid1/slrl* double mutants, genetic crosses were performed between heterozygous plants for each gene, and the genotype of each F₂ plant was identified by the sequence of the genes. All rice plants were grown in a greenhouse at 30 °C (day) and 24 °C (night).

GA-responsive experiments. GA-induced elongation of second leaf sheath and α-amylase induction in embryoless half seeds were performed as described previously²⁶.

Plasmid construction. Full-length *GID1* cDNAs were produced by RT–PCR from total RNA from wild-type and *gid1* mutant alleles. For production of recombinant GST–GID1 protein, the *GID1* cDNA sequences were inserted into pGEX-4T (Pharmacia). For complementation, the rice genomic DNA from a bacterial artificial chromosome (BAC) clone was digested with *Pst*I, and a 6.7-kb DNA fragment that included the entire *GID1* sequence was cloned into a pBluescript vector. The *Pst*I fragment was blunt-end-filled and inserted into the *Sma*I site of the hygromycin-resistant binary vector pGI-Hm12 (provided by H. Hirano). Act1 promoter–GID1 and Act1 promoter–GID1–GFP constructs were produced by PCR from GID1 and GID–GFP cDNA and inserted into the pBActInos vector²⁹. The construction of SLR1 promoter–SLR1–GFP⁵ and the introduction of constructs into rice cells by *Agrobacterium tumefaciens*-mediated transformation were performed as previously described³⁰. Control plants for GID1 overexpressors were produced by transformation of pBActInos vector.

GA-binding assay. Recombinant GST–GID1 and its mutant proteins for the *in vitro* GA-binding assay were expressed in *Escherichia coli* and purified using glutathione beads according to standard protocols. The amount of purified recombinant proteins was quantified by the Bio-Rad Protein Assay system. For

the GA-binding assay, [1,2,16,17-³H₄]16,17-dihydro-GA₄ was used as a labelled GA₄. This labelled GA₄ was synthesized with the help of Du Pont/NEN. *In vitro* assays for GA binding were performed as described previously¹⁴ with some modifications. Purified GST–GID1 proteins (16 µg) were dissolved in 300 µl binding buffer (20 mM Tris–HCl (pH 7.6), 5 mM 2-mercaptoethanol and 0.1 M NaCl) and incubated with 100 µl ³H₄-16,17-dihydro-GA₄ (6 pmol), either with an 833-fold excess of unlabelled GA₄ for nonspecific binding or without excess unlabelled GA₄ for total binding. Afterwards, 100 µl of the mixture was fractionated on a NAP-5 column (Amersham Biosciences). After discarding a void volume binding buffer eluate (600 µl), a 200-µl fraction was collected and its radioactivity measured. The specific binding activity, which reflected the number of replaceable GA-binding sites, was calculated by subtraction of nonspecific binding from total binding. To examine the saturability of binding of GST–GID1 to 16,17-dihydro-GA₄, GST–GID1 was incubated with 100 µl ³H₄-16,17-dihydro-GA₄ (6 pmol) and increasing concentrations of unlabelled 16,17-dihydro-GA₄. From the radioactivity of the NAP-5 fraction and the labelled/unlabelled ratio in the assay mixture, total binding of 16,17-dihydro-GA₄ (labelled 16,17-dihydro-GA₄ plus unlabelled 16,17-dihydro-GA₄) was calculated.

Yeast two-hybrid assay. The Matchmaker Two-Hybrid System (Clontech) was used for the yeast two-hybrid assay. pGBKT7–GID1 served as the bait and pGADT7–SLR1 as the prey. Plate assays (–His) and β-galactosidase liquid assays were performed according to the manufacturer's protocol, with the modification that the plate and liquid media either contained 10^{–4} M GA₃ or not.

Measurements of endogenous GAs and other analyses. For the measurement of endogenous GAs, we collected 1-month-old rice shoots. Purification and the quantitative analysis of endogenous GAs by gas chromatography–selected ion monitoring were described previously³¹. RNA gel blot analysis and western blot analysis were performed as described elsewhere³.

Received 10 May; accepted 12 July 2005.

- Davies, P. J. *Plant Hormones* (Kluwer, Dordrecht, The Netherlands, 1995).
- Hedden, P. & Phillips, A. L. Gibberellin metabolism. New insights revealed by the genes. *Trends Plant Sci.* 5, 523–530 (2000).
- Sakamoto, T. *et al.* An overview of gibberellin metabolism enzyme genes and their related mutants in rice. *Plant Physiol.* 134, 1642–1653 (2004).
- Ikeda, A. *et al.* *Slender rice*, a constitutive gibberellin response mutant, is caused by a null mutation of the *SLR1* gene, an ortholog of the height-regulating gene *GAI/RGA/RHT/D8*. *Plant Cell* 13, 999–1010 (2001).
- Itoh, H., Ueguchi-Tanaka, M., Sato, Y., Ashikari, M. & Matsuoka, M. The gibberellin signalling pathway is regulated by the appearance and disappearance of SLENDER RICE1 in nuclei. *Plant Cell* 14, 57–70 (2002).
- Peng, J. *et al.* The *Arabidopsis* *GAI* gene defines a signalling pathway that negatively regulates gibberellin responses. *Genes Dev.* 11, 3194–3205 (1997).
- Silverstone, A. L., Ciampaglio, C. N. & Sun, T.-P. The *Arabidopsis* *RGA* gene encodes a transcriptional regulator repressing the gibberellin signal transduction pathway. *Plant Cell* 2, 155–169 (1998).
- Gubler, F., Chandler, P. M., White, R. G., Llewellyn, D. J. & Jacobsen, J. V. Gibberellin signalling in barley aleurone cells. Control of SLR1 and GAMYB expression. *Plant Physiol.* 129, 191–200 (2002).
- Itoh, H., Matsuoka, M. & Steber, C. M. A role for the ubiquitin-26S-proteasome pathway in gibberellin signalling. *Trends Plant Sci.* 8, 492–497 (2003).
- Sasaki, A. *et al.* Accumulation of phosphorylated repressor for gibberellin signalling in an F-box mutant. *Science* 299, 1896–1898 (2003).
- Gomi, K. *et al.* GID2, an F-box subunit of the SCF E3 complex, specifically interacts with phosphorylated SLR1 protein and regulates the gibberellin-dependent degradation of SLR1 in rice. *Plant J.* 37, 626–634 (2004).
- Hooley, R. *et al.* Gibberellin perception and the Avena Fatua aleurone: do our molecular keys fit the correct locks? *Biochem. Soc. Trans.* 20, 85–89 (1992).
- Lovegrove, A., Barratt, D. H., Beale, M. H. & Hooley, R. Gibberellin-photoaffinity labelling of two polypeptides in plant plasma membranes. *Plant J.* 15, 311–320 (1998).
- Nakajima, M. *et al.* Partial purification and characterization of a gibberellin-binding protein from seedlings of *Azuki angularis*. *Biochem. Biophys. Res. Commun.* 241, 782–786 (1997).
- Itoh, H., Ueguchi-Tanaka, M., Sentoku, N., Kitano, H. & Matsuoka, M. Cloning and functional analysis of two gibberellin 3β-hydroxylase genes that are differently expressed during the growth of rice. *Proc. Natl Acad. Sci. USA* 98, 8909–8914 (2001).
- Thornton, T. M., Swain, S. M. & Olszewski, N. E. Gibberellin signal transduction presents ellipsis the SPY who O-GlcNAc'd me. *Trends Plant Sci.* 4, 424–428 (1999).
- Sasaki, A. *et al.* Green revolution: a mutant gibberellin-synthesis gene in rice. *Nature* 416, 701–702 (2002).
- Itoh, H. *et al.* Dissection of the phosphorylation of rice DELLA protein, SLENDER RICE1. *Plant Cell Physiol.* 46, 1392–1399 (2005).
- Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.* 33, 192–196 (2005).

20. Osterlund, T. *et al.* Domain-structure analysis of recombinant rat hormone-sensitive lipase. *Biochem. J.* **319**, 411–420 (1996).
21. Manco, G. *et al.* Cloning, overexpression, and properties of a new thermophilic and thermostable esterase with sequence similarity to hormone-sensitive lipase subfamily from the archaeon *Arcaeglobus fulgidus*. *Arch. Biochem. Biophys.* **373**, 182–192 (2000).
22. Osterlund, T. Structure-function relationships of hormone-sensitive lipase. *Eur. J. Biochem.* **268**, 1899–1907 (2001).
23. Nishijima, T., Koshioka, M. & Yamazaki, H. Use of several gibberellin biosynthesis inhibitors in sensitized rice seedling bioassays. *Biosci. Biotech. Biochem.* **58**, 572–573 (1994).
24. Nakayama, I. *et al.* Effects of a new plant growth regulator prohexadine calcium (BX-112) on shoot elongation caused by exogenously applied gibberellins in rice (*Oryza sativa* L.) seedlings. *Plant Cell Physiol.* **31**, 195–200 (1990).
25. Natsume, T., Hirota, J., Yoshikawa, F., Furuichi, T. & Mikoshiba, K. Real time analysis of interaction between inositol 1,4,5-triphosphatereceptor type I and its ligand. *Biochem. Biophys. Res. Commun.* **260**, 527–533 (1999).
26. Ueguchi-Tanaka, M. *et al.* Rice dwarf mutant *d1*, which is defective in the α subunit of the heterotrimeric G protein, affects gibberellin signal transduction. *Proc. Natl Acad. Sci. USA* **97**, 11638–11643 (2000).
27. Dharmasiri, N., Dharmasiri, S. & Estelle, M. The F-box protein TIR1 is an auxin receptor. *Nature* **435**, 441–445 (2005).
28. Keinski, S. & Leyser, O. The *Arabidopsis* F-box protein TIR1 is an auxin receptor. *Nature* **435**, 446–451 (2005).
29. Hiei, Y., Ohta, S., Komari, T. & Kumashiro, T. Efficient transformation of rice (*Oryza Sativa* L.) mediated by *Agrobacterium* and sequence analysis of the boundaries of the T-DNA. *Plant J.* **6**, 270–282 (1994).
30. Nishimura, A., Ito, M., Kamiya, N., Sato, Y. & Matsuoka, M. *OsPNH1* regulates leaf development and maintenance of the shoot apical meristem in rice. *Plant J.* **30**, 189–201 (2002).
31. Kobayashi, M. *et al.* Fluctuation of endogenous gibberellin and abscisic acid levels in the germinating seeds of barley. *Biosci. Biotechnol. Biochem.* **59**, 1969–1970 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Oomiya, S. Hattori and I. Aichi for technical assistance, and C. Ueguchi for suggestions regarding the yeast two-hybrid assay. This work was supported in part by a Grant-in-Aid for the Center of Excellence, the Program for the Promotion of Basic Research Activities for Innovative Bioscience (M.M. and H.K.), the MAFF Rice Genome Project, IP1003 (M.A. and M.M.), and by the Ministry of Education, Culture, Sports, Science and Technology of Japan (I.Y., M.N. and M.U.-T.).

Author Information Sequence data from this article have been deposited in the DDBJ/EMBL/GenBank databases under accession number AB211399. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.M. (makoto@nuagr1.agr.nagoya-u.ac.jp).

Regulated cell-to-cell variation in a cell-fate decision system

Alejandro Colman-Lerner^{1*}, Andrew Gordon^{1*}, Eduard Serra¹, Tina Chin¹, Orna Resnekov¹, Drew Endy², C. Gustavo Pesce¹ & Roger Brent¹

Here we studied the quantitative behaviour and cell-to-cell variability of a prototypical eukaryotic cell-fate decision system, the mating pheromone response pathway in yeast. We dissected and measured sources of variation in system output, analysing thousands of individual, genetically identical cells. Only a small proportion of total cell-to-cell variation is caused by random fluctuations in gene transcription and translation during the response ('expression noise'). Instead, variation is dominated by differences in the capacity of individual cells to transmit signals through the pathway ('pathway capacity') and to express proteins from genes ('expression capacity'). Cells with high expression capacity express proteins at a higher rate and increase in volume more rapidly. Our results identify two mechanisms that regulate cell-to-cell variation in pathway capacity. First, the MAP kinase Fus3 suppresses variation at high pheromone levels, while the MAP kinase Kss1 enhances variation at low pheromone levels. Second, pathway capacity and expression capacity are negatively correlated, suggesting a compensatory mechanism that allows cells to respond more precisely to pheromone in the presence of a large variation in expression capacity.

Biological systems are composed of physical constituents that constrain their performance. However, some aspects of system performance, including cell-to-cell variation, are often regulated by active mechanisms^{1–8}. The study of variation in the behaviour of genetically identical cells goes back as far as Delbrück⁹, who measured differences in the numbers of phage T1 produced by individual, singly infected *E. coli*.

Recently, a number of studies have used fluorescent protein reporters to study cell-to-cell variation in gene expression^{10–16}. For example, variation in gene expression among genetically identical bacteria has been studied by measuring the correlation in expression of two different fluorescent protein reporter genes under control of the same promoters¹¹. Cell-to-cell variation resulted from both stochastic fluctuations in the expression of each reporter protein (termed 'intrinsic noise') and differences in the levels of cellular components needed for expression of both reporters (termed 'extrinsic noise'), and the results suggested that some components of extrinsic noise affect gene expression in general¹¹. The component of extrinsic noise that affects overall gene expression has recently been quantified in *Escherichia coli*^{12,13}. This type of gene expression analysis has also been performed in yeast¹⁴, and revealed that intrinsic noise contributes little to cell-to-cell variation in gene expression. Cell-to-cell variation in the expression of two non-identical promoters was correlated, consistent again with the idea that some extrinsic noise is a result of global differences in gene expression. Others have shown that changes in the amount of transcription and translation affect the amount of overall cell-to-cell variation in the expression of a single reporter^{15,16}.

Here we studied cell-to-cell variation, not in gene expression, but in the quantitative output of a cell-fate decision system: the pheromone response pathway in the yeast *Saccharomyces cerevisiae*. In haploid cells of the α mating type, α -factor (a pheromone secreted by cells of the α mating type) triggers a fate decision to switch from

normal, vegetative growth to the initiation of mating events, including induction of gene transcription, cell cycle arrest and changes in morphology. The pathway is a prototypical eukaryotic signal transduction system that includes a G-protein-coupled receptor and a MAP kinase cascade¹⁷ (Fig. 1a).

To study cell-to-cell variation in the workings of this decision system, we used pheromone-induced expression of fluorescent protein reporter genes as a readout. We realized that cell-to-cell differences in the levels of fluorescent proteins would convolve differences in the operation of the signal transduction pathway with cell-to-cell differences in gene expression from the reporters. To distinguish between and quantify these two contributions, we generated a series of yeast strains containing genes for yellow and cyan fluorescent protein (YFP and CFP). We compared the results from experiments in which YFP and CFP were controlled by identical α -factor-responsive promoters with results from experiments in which YFP was driven by an α -factor-responsive promoter and CFP by an α -factor-independent promoter (Fig. 1b, c).

We constructed an analytical framework to guide the design and interpretation of these experiments. We considered the α -factor response pathway and the means used to measure its activity (reporter gene expression) as a single system composed of two connected subsystems: 'pathway' and 'expression' (Fig. 1a). In each subsystem, we distinguished two sources of variation: stochastic fluctuations and cell-to-cell differences in 'capacity'. Capacity depends on the number, localization and activity of proteins that transmit the signal (pathway capacity) or express genes into proteins (expression capacity), and is determined by the state of the cells at the start of the experiment. We limited the term 'noise' to refer to the variation due to stochastic fluctuations in subsystem function that occur during the experiment (for example, spontaneous differences in the occurrence and timing of discrete probabilistic chemical reactions). By distinguishing these two sources, we modify the

¹The Molecular Sciences Institute, 2168 Shattuck Avenue, Berkeley, California 94704, USA. ²Division of Biological Engineering, Massachusetts Institute of Technology, 31 Ames Street, Building 68-580, Cambridge, Massachusetts 02139, USA.

*These authors contributed equally to this work.

terminology used in ref. 11, which used the term noise to refer to both sources of variation.

If we could stimulate a cell numerous times by going back in time and repeating the experiment (thereby guaranteeing the same initial state), the average system output for all the repeated trials would be a measure of the expectation value of the output. According to our framework, this value would depend on pathway capacity and expression capacity. Any differences in system output on individual trials would arise from random fluctuations during each trial in the number of molecules and the workings of the machineries (1) transmitting the signal (transmission noise) and (2) transcribing the reporter messenger RNA and translating it into protein (expression noise). If we performed the above thought experiment

on a different cell, a different average value for system output might be obtained. This cell-to-cell difference in average system output would be the manifestation of a cell-to-cell difference in the capacities of the two subsystems—caused, for example, by a pre-existing difference in the number of molecules that transmit the signal or express proteins from genes.

Analytical framework

The first subsystem, the α -factor response pathway, includes all steps that lead to activation of transcription, which depends on the activity of DNA-bound transcription factors. The input to the pheromone pathway is α -factor and the output depends on the total amount of active transcription factor (Ste12) bound upstream of the α -factor-

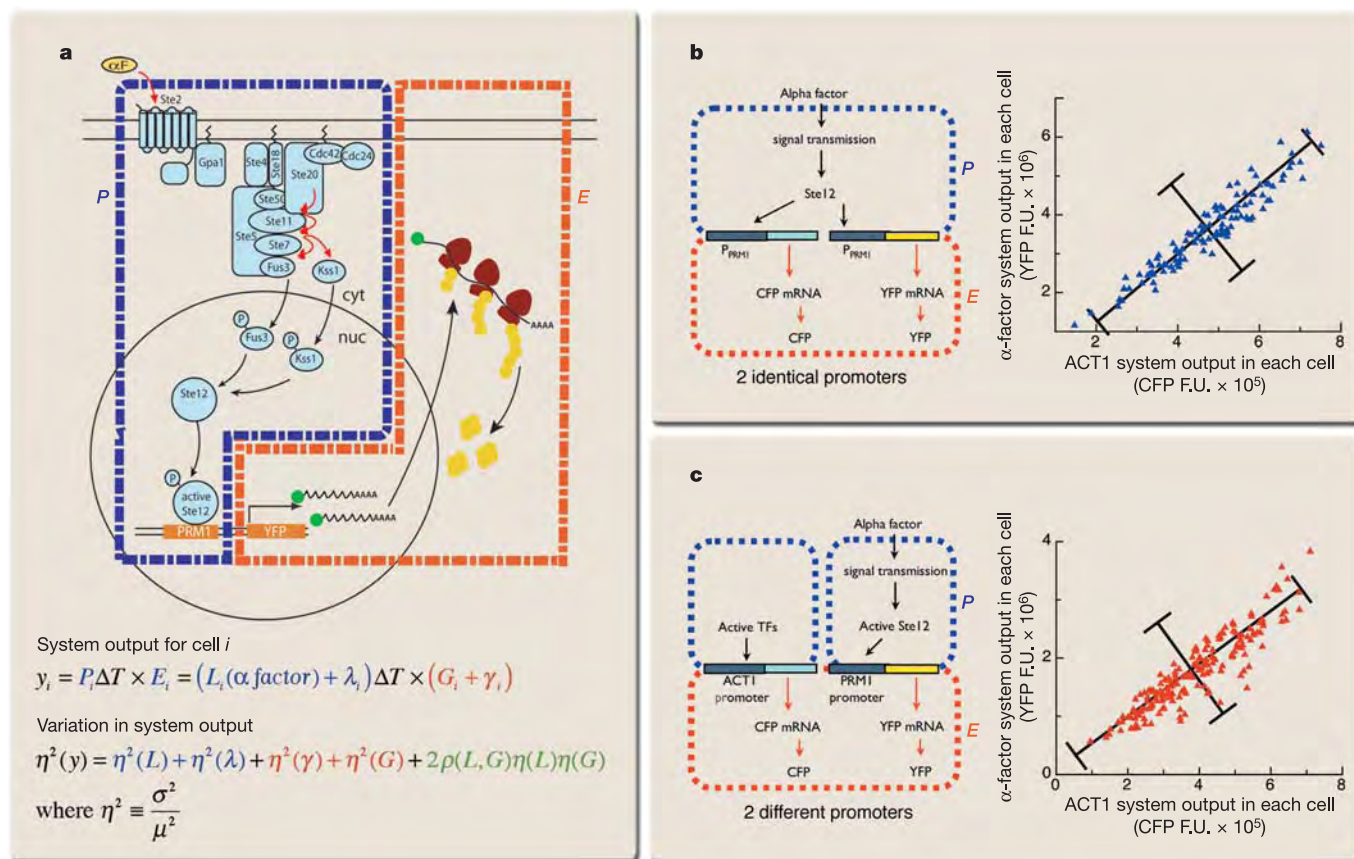


Figure 1 | Quantifying sources of cell-to-cell variation. **a**, The mating pheromone response system and analytical framework for decomposition. Diagram shows proteins in the yeast cell membrane, cytoplasm (cyt), and nucleus (nuc). Events in the blue box are classified as the pathway subsystem. The binding of α -factor to the receptor Ste2 causes dissociation of the heterotrimeric G protein α -subunit Gpa1 from the Ste4–Ste18 dimer ($\beta\gamma$ -subunits). Ste4 recruits the scaffold protein Ste5 to the membrane, and Ste5 binds the MEKK Ste11, the MEK Ste7 and the MAPK Fus3. The PAK kinase Ste20 initiates the MAPK cascade by activating Ste11, which activates Fus3, which in turn activates the MAP kinases Fus3 and Kss1. Phosphorylated Fus3 and Kss1 leave Ste5 and translocate to the nucleus, where they activate the transcription factor Ste12. At a given concentration of α -factor, the amount of activated Ste12 on the promoter is the ‘pathway subsystem output’ P (defined in the text). Events in the red box are classified as the expression subsystem, quantified by E (defined in the text). E includes transcription initiation, mRNA elongation and processing, nuclear export and cytoplasmic protein translation. The total system output—the amount of fluorescent reporter protein y produced in any cell i —depends on P , E , α -factor concentrations and the duration of stimulation ΔT . To measure cell-to-cell variation in the population we used the normalized variance η^2 , which is decomposed into separate additive terms that represent different sources of cell-to-cell variation as described in the text. **b**, Type I experiment, measuring gene expression noise (γ). In strains containing two identical

α -factor-responsive promoters driving the YFP and CFP reporter genes, the same pathway (blue box) and expression machinery (red box) controls the production of reporter proteins. We stimulated TCY3096 cells with a high concentration (20 nM) of α -factor and collected YFP and CFP images after 3 h. Each cell is represented by a single symbol showing its YFP and CFP signals (in F.U. or fluorescent units). The uncorrelated variation between YFP and CFP can be seen as the width of the minor axis, which is orthogonal to the 45° diagonal major axis (lines in black); it is caused only by stochastic variation in gene expression (γ). We used the orthogonal scatter as a measure of $\eta^2(\gamma)$, here 0.002 ± 0.0001 . See Table 1 and Supplementary Fig. S4 for the gene expression noise shown by other promoters and other α -factor concentrations. **c**, Type II experiment, measuring variation in pathway subsystem output (P) and expression capacity (E). In strains containing different promoters driving the YFP and CFP reporter genes, different subsystems (blue boxes) regulate the activity of the DNA-bound transcription factors, but the subsystem enabling expression of the reporter genes (red box) is the same. We stimulated TCY3154 cells as in the type I experiment above. Variation in expression capacity affected only the correlated variation (the dispersion of points along the major axis, or the 45° diagonal). The uncorrelated variation (the dispersion of points along the minor axis) is due to the gene expression noise measured from type I experiments and to cell-to-cell variations in the pathway subsystems for each promoter.

responsive reporter gene, for the time elapsed since addition of input, ΔT . For any individual cell i , the pathway output is $P_i \Delta T$. The term P_i is the average output per unit time, which is the sum of L_i , the expectation value of this average (which we call pathway power) and λ_i , the stochastic fluctuations in P_i . Pathway power L is a function of the input and pathway capacity. Note that the same analysis can apply to any pathway that leads to activation of transcription factors bound upstream of genes (including constitutive or 'housekeeping' genes) that one might not normally think of as responsive to signals.

The second subsystem, reporter gene expression, includes all events from transcription initiation through to the accumulation of protein. For α -factor reporter genes, the input to this subsystem is the output from the α -factor pathway subsystem, and the output of the expression subsystem is the amount of mature fluorescent reporter protein. For any individual cell i , expression output per unit of promoter activity (E_i) is the sum of G_i , the expectation value of E_i , and γ_i , the stochastic fluctuations of E_i . G_i measures the ability

of cell i to express the reporter protein from the gene; it is independent of the input level (see below) and is determined primarily by expression capacity, a global cellular property that equally affects expression of other genes. From here on, we will refer to G as expression capacity. In contrast, the value of γ refers only to the stochastic fluctuations in the levels of reporter protein.

We assumed that induction of reporters (and the other genes induced by α -factor¹⁸) did not significantly decrease cell-wide gene expression capacity. Two facts supported this assumption. First, at a given stimulus (concentration and treatment duration of α -factor), cells with two or three copies of the α -factor-inducible reporter produced a corresponding two- or threefold increase in fluorescent protein levels (data not shown). Second, after α -factor treatment, expression of fluorescent proteins controlled by constitutive promoters was unchanged (Supplementary Fig. S1 and data not shown). These findings suggest that the overall working of the expression subsystem is independent of the pheromone pathway subsystem.

We then described system output, the amount of reporter protein y in cell i , as the product of P_i , the average pathway subsystem output per unit time, ΔT , the time since addition of α -factor, and E_i , the expression output per unit of pathway subsystem output:

$$y_i = P_i \Delta T \times E_i \quad (1)$$

where $P_i = L_i + \lambda_i$ and $E_i = G_i + \gamma_i$. For α -factor response, P might vary with E , but E would not vary with P . For example, a higher E might increase (or decrease) P if a higher E leads to an increased (or decreased) ratio of positive regulators to negative regulators of the pathway. This potential dependency generates the correlation term in equation (2) below.

We defined variation (η^2) in system output among cells as the 'normalized variance', the variance (σ^2) divided by the mean squared (μ^2) (that is, $\eta^2 = \sigma^2/\mu^2$). As derived in Supplementary Materials, the variation in y for a population of cells is described by the sum of the individual sources of variation, plus a correlation term:

$$\eta^2(y) = \overbrace{\eta^2(L) + \eta^2(\lambda)}^{\eta^2(P)} + \overbrace{\eta^2(\gamma) + \eta^2(G)}^{\eta^2(E)} + 2\rho(L, G)\eta(L)\eta(G) \quad (2)$$

where $\eta^2(L)$ is the variation in pathway power, $\eta^2(\lambda)$ is transmission noise, $\eta^2(G)$ is variation in the expression capacity, $\rho(L, G)$ is the correlation coefficient between L and G , and $\eta^2(\gamma)$ is expression noise. This last term is equivalent to 'intrinsic noise', and total variation $\eta^2(y)$ minus $\eta^2(\gamma)$ is equivalent to 'extrinsic noise', as used in ref. 11. The term $\eta^2(P)$ is the variation in average pathway subsystem output per unit time, which, because pathway subsystem output is given by $P\Delta T$ and ΔT is the same for every cell, is equivalent to variation in pathway subsystem output. The term $2\rho(L, G)\eta(L)\eta(G)$ accounts for a possible correlation between expression capacity and pathway power, as discussed above. This term increases or decreases the total variation depending on the sign of the correlation coefficient. We assumed here that the pathway and expression subsystems do not share molecular components, and therefore that stochastic fluctuations in the expression of the reporter protein (γ) are not correlated with the pathway subsystem (L or λ); similarly, λ is not correlated to the expression subsystem (G or γ). Thus, λ and γ do not appear in any correlation terms.

From analytical framework to experimental design

To measure the contributions of the different sources of cell-to-cell variation in system output, we constructed haploid yeast strains that contained different combinations of two promoters driving the expression of YFP and CFP. We used strains with YFP and CFP reporters driven by the α -factor-responsive P_{PRM1} promoter¹⁹ or by the α -factor-independent P_{ACT1} promoter (Fig. 1 and Supplementary Table S1). We treated cells attached to the glass bottom of microtitre wells with α -factor and captured images at intervals using an inverted epifluorescence microscope and a CCD camera. Custom

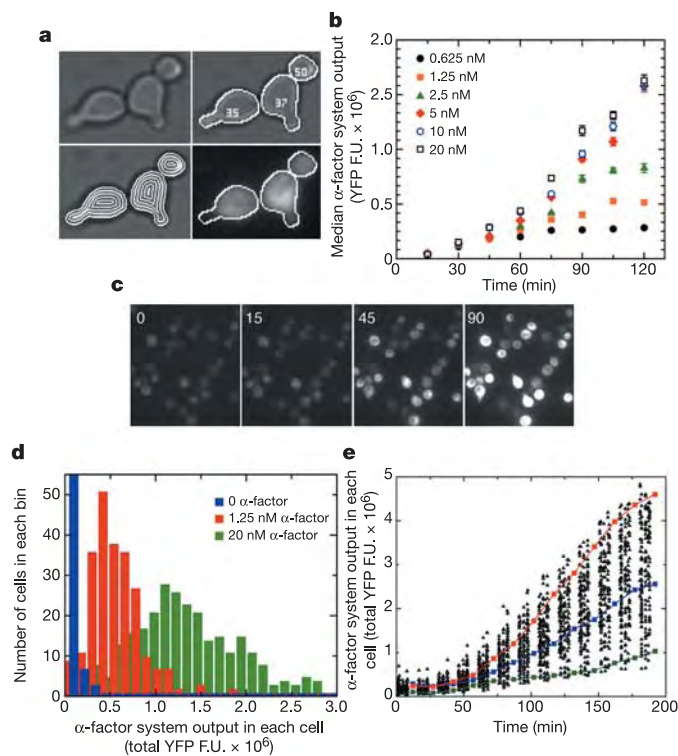


Figure 2 | Quantification of system output in single cells during α -factor response. We treated ACLY387 (P_{PRM1} -YFP) cells with the indicated concentrations of α -factor at time zero, acquired YFP images every 15 min for 3 h and analysed them using Cell-ID. **a**, A sampling of cells exposed to 100 nM α -factor for 75 min. For a bright-field image (top left), we used Cell-ID to locate, determine the perimeter of, and number (top right) each cell, then draw consecutively tighter annuli (bottom left, white lines show every other annulus) to calculate cell volume. The boundary contour of the cells was then transferred to the corresponding fluorescence image (bottom right) and the fluorescence intensity of the enclosed pixels was summed. **b**, Time-dependent dose-response. Data correspond to the median system output $\pm$ s.e.m.; $n = 400$ –600 cells. **c**, YFP fluorescence images of cells at the indicated times (in min), showing cell-to-cell variation in system output. **d**, Distribution of system output in populations of yeast exposed to α -factor for 2 h. **e**, System output of individual cells treated with 20 nM α -factor. We tracked three fields of cells treated identically. Images of each field were captured every 15 min, generating three 'columns' of cells at each interval. Each triangle represents the output of a single cell at a single time point. The trajectories of a representative strong (red), medium (blue) and weak (green) responding cell are shown, connected by lines and shifted rightward by 8 min to aid visualization. System output is measured in fluorescent units (F.U.), with one F.U. corresponding to approximately 2.5 photons hitting the CCD chip.

image-analysis software (Cell-ID 1.0) was used to extract measurements of individual cells from the images and to correct the measured fluorescence intensity to account for photobleaching and for the fact that cells of different size have a different fraction of their volume in focus (Fig. 2a and A.G., A.C.-L., T.C., K. Benjamin & R.B., submitted manuscript). We measured YFP and CFP fluorescence in large numbers of genetically identical cells treated with uniform amounts of α -factor, and computed the cell-to-cell variation in fluorescence. Despite differences in size, these cells each had one YFP and one CFP gene in their genome, and so to measure reporter gene activity we calculated total fluorescence per cell rather than fluorescence per unit volume. (In Fig. 3c and Supplementary Fig. S2, we examined the relationship between total reporter protein per cell and cell volume.)

We performed two types of experiments. In type I experiments, we used cells with identical promoters driving the expression of YFP and CFP (Fig. 1b). Although both constructs were controlled by the same pathway and shared the same expression subsystem, they were two separate genes. Therefore, differences in the levels of CFP and YFP in each cell should be due to expression noise $\eta^2(\gamma)$ (ref. 11) (and also to fluctuations in the amount of active transcription factor at each promoter; see Supplementary Materials section 5.3) and not to transmission noise $\eta^2(\lambda)$. In type II experiments, we used cells with different promoters driving the expression of CFP and YFP (Fig. 1c). These experiments differed from Type I experiments in that the two promoters were now controlled by different and independent pathways. Therefore, differences in the levels of CFP and YFP in each cell should be due not only to expression noise but also to transmission noise and differences in promoter-specific pathway subsystem power. In both types of experiments, the CFP and YFP genes shared the same expression subsystem. Therefore, cell-to-cell variation in expression capacity G only caused cell-to-cell differences in CFP and YFP levels, not differences within a given cell (Fig. 1c).

Mathematical analysis of measurements from the two types of experiments allowed us to quantify variation in expression capacity $\eta^2(G)$, expression noise $\eta^2(\gamma)$, variation in pathway subsystem output $\eta^2(P)$ (a measure that combined $\eta^2(L)$ and $\eta^2(\lambda)$, which we have not separated experimentally), and the covariance term $2\rho(L,G)\eta(G)\eta(L)$ (see Supplementary Materials), and it also suggested approaches for examining mechanisms that might regulate the different sources of variation.

Large variation in expression capacity

In cells exposed to high concentrations of α -factor (20 nM), we detected induced fluorescence within the first 30 min (Fig. 2b). Much

of this delay was caused by the slow maturation of the YFP and CFP fluorophores ($T_{1/2} = 39$ and 49 min for YFP and CFP, respectively) (A.G., A.C.-L., T.C., K. Benjamin & R.B., submitted manuscript). YFP and CFP mature at slightly different rates, but this did not affect the results presented below (see Supplementary Materials, section 3). The measurements were sensitive to low pathway activation—we readily detected output in single cells stimulated with 0.1 nM α -factor, a concentration 30 times lower than that needed for half-maximal output (3 nM, which is a good match to the published K_d of α -factor for its receptor²⁰). Cells differed greatly in system output; the top 5% of cells showed approximately fourfold higher output than the bottom 5% of cells. The distribution of system output in the population was roughly bell-shaped at all concentrations tested (Fig. 2d and not shown), indicating that system output as measured by our reporter shows a graded (as opposed to an all-or-nothing) response to α -factor (see also Supplementary Fig. S3).

Variation in α -factor system output (η^2) was relatively constant over time (Fig. 3a). This suggests that most of the variation was due to cell-to-cell differences already present at the time of addition of α -factor. If the observed variation were caused by accumulation of signal transmission noise or gene expression noise, the standard deviation should grow as the square root of the mean (as expected for Poisson processes) rather than linearly, as our data demonstrated (Fig. 3b). Consistent with this interpretation, only a small proportion ($1.2 \pm 0.1\%$) of the observed variation was caused by gene expression noise. This was shown by the narrow dispersion of the data points along the minor axis in Fig. 1b, which shows YFP and CFP levels derived from genes controlled by identical P_{PRM1} promoters. We also observed low gene expression noise for several other, α -factor-independent reporters (Supplementary Table S3 and Supplementary Fig. S4). Notably, variation in α -factor pathway subsystem output ($\eta^2(L) + \eta^2(\lambda)$) was also small (17% of the total) in cells stimulated with high concentrations of α -factor (20 nM). This was derived from the dispersion of the points along the minor axis of Fig. 1c, where YFP and CFP levels from the P_{PRM1} and P_{ACT1} promoters are shown, respectively. The bulk of the total variation ($\eta^2 = 0.17 \pm 0.02$, 3 h after addition of 20 nM α -factor) was caused by cell-to-cell differences in expression capacity ($\eta^2(G) = 0.14 \pm 0.02$), as shown by the wider dispersion of data points along the major compared to the minor axis in Fig. 1c ($\rho_{YFP,CFP} = 0.88 \pm 0.05$). Details of these calculations are provided in the Supplementary Information.

We obtained consistent results using two other constitutive promoters (P_{STE5} and P_{BMH2}) to drive CFP expression (data not

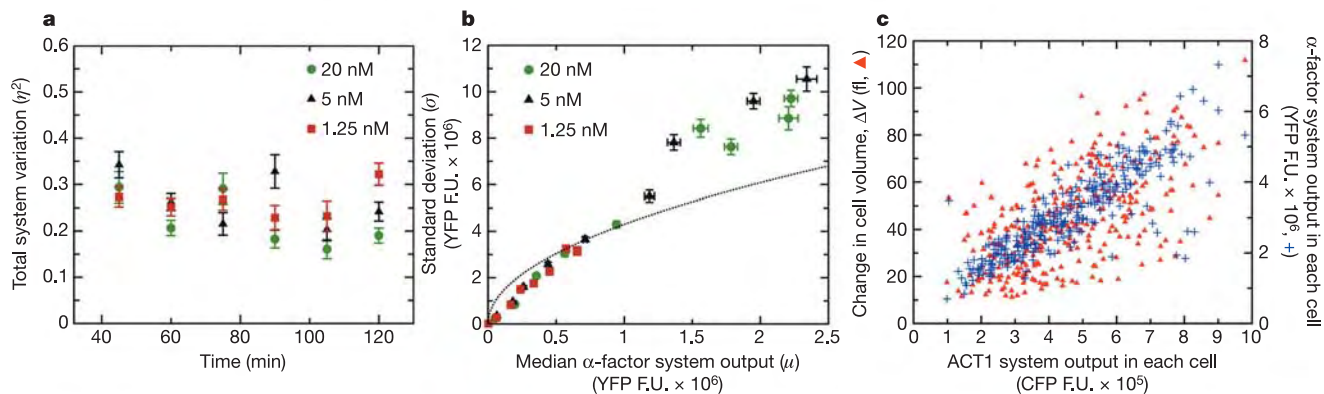


Figure 3 | Cell-to-cell variation is dominated by initial differences between cells. We treated ACLY387 cells (P_{PRM1} -YFP) with α -factor as in Fig. 2 and measured cell-to-cell variation. **a**, Total variation $\eta^2(y)$ over time after addition of α -factor. **b**, Standard deviation σ was plotted against median system output $\mu \pm$ s.e.m. for populations of cells treated with the indicated concentrations of α -factor for different times. The dotted line corresponds to a square root function of μ (forced to pass through the point $x = 1 \times 10^6$,

$y = 4 \times 10^6$), the relationship expected if most of the variation in the population originated by stochastic processes over the course of the experiment. **c**, We treated TCY3154 cells (P_{PRM1} -YFP, P_{ACT1} -CFP) with 20 nM α -factor for 3 h. Data correspond to single-cell values of ACT1 system output (total CFP at 3 h minus total CFP at time zero) versus the change in volume, ΔV (left y axis, red, $\rho_{CFP,\Delta V} = 0.77$), and versus α -factor system output (right y axis, blue, $\rho_{CFP,YFP} = 0.89$).

shown). Moreover, we observed a strong correlation between YFP and CFP in strains with reporters controlled by two different α -factor-independent promoters (Supplementary Fig. S5 and data not shown). The above results suggest that in yeast, cell-to-cell variation in gene expression is dominated by differences in expression capacity among cells, and that expression capacity is a global cellular feature controlling expression of many or most genes.

As cells with high expression capacity produce protein more rapidly, we reasoned that cells with high expression capacity might increase faster in volume (larger ΔV). To test this, we examined the correlation between cell volume, the rate of change in cell volume and the expression of fluorescent reporters. There was a low correlation between P_{ACT1} -derived CFP and initial volume (Supplementary Fig. S2). However, we observed a significant correlation between

P_{ACT1} -derived CFP and ΔV ($\rho = 0.77$, Fig. 3c), suggesting that expression capacity might help to determine the rate of increase in cell volume. However, the fact that the correlation of P_{ACT1} -CFP with P_{PRM1} -YFP ($\rho = 0.89$) was significantly better than the correlation of P_{ACT1} -CFP with ΔV indicates that other factors, uncorrelated with expression capacity, also influenced ΔV .

Effect of cell cycle on variation

One conspicuous difference among exponentially growing yeast is their cell-cycle position. To determine the effects of cell-cycle position on cell-to-cell variation in system output, we measured variation in system output in yeast arrested from cycling. We replaced the cyclin-dependent kinase Cdc28 with an engineered variant (Cdc28-as2)²¹ that is sensitive to the chemical inhibitor 1-NM-PP1 (4-amino-1-(tert-butyl)-3-(1'-naphthylmethyl)pyrazolo[3,4-D]pyrimidine). In the absence of inhibitor, Cdc28-as2 cells behave the same as wild-type cells (not shown), but addition of 10 μ M inhibitor arrests Cdc28-as2 cells at the G2/M transition²¹. We stimulated cells containing Cdc28-as2 and the P_{PRM1} -YFP and P_{ACT1} -CFP reporters (TCY3154 cells) with 20 nM α -factor with or without inhibitor, and followed them over time. Visual examination revealed that without inhibitor, cells with small buds or those about to initiate bud formation showed a delay in pathway induction (Fig. 4a), consistent with reports that cells at the G1/S transition cannot respond to pheromone because a cyclin-dependent kinase complex inhibits the MAP kinase cascade^{22–24}. With inhibitor, cells began to induce the pathway at almost the same time (Fig. 4a), reducing total α -factor system output variation by 45% (from $\eta^2(y) = 0.19 \pm 0.03$ to $\eta^2(y) = 0.11 \pm 0.01$, Fig. 4b). Addition of inhibitor did not alter cell-to-cell variation in P_{ACT1} -CFP signal (not shown), indicating that inhibition of Cdc28 reduced variation specific to the α -factor pathway subsystem and not variation in the gene expression subsystem. We obtained similar results by synchronizing the cell cycle in late mitosis using the Cdc15-2ts mutant (not shown), suggesting that all the measured cell-cycle-dependent pathway variation was due to Cdc28 activity.

In addition to eliminating a source of variation, inhibiting Cdc28 allowed us to study system output at low concentrations of α -factor. Only concentrations of α -factor above 2.5 nM caused uniform cell-cycle arrest. At lower concentrations, cells that continued to divide diluted the reporter protein into daughter cells, complicating the measurement of cell-to-cell variation in system output. However, in the presence of Cdc28-as2 inhibitor, all cells arrest.

Large pathway power variation at low α -factor concentrations

The above experiments determined that at high concentrations of α -factor (>20 nM), less than 25% of cell-to-cell variation in system output was due to differences in pathway subsystem output, and more than 75% was due to differences in expression capacity (Table 1). To determine whether the relative contributions of the sources of variation in system output depended on the concentration of α -factor, we stimulated TCY3154 cells with lower α -factor concentrations in the presence of Cdc28 inhibitor. At low concentrations, we observed a reduced correlation between P_{PRM1} -YFP and P_{ACT1} -CFP ($\rho_{YFP,CFP} = 0.94 \pm 0.01$ at 20 nM and $\rho_{YFP,CFP} = 0.72 \pm 0.02$ at 1.25 nM, Fig. 5a). A control experiment demonstrated that α -factor concentration did not affect correlation between two α -factor-independent promoters (Supplementary Fig. S5). At low concentrations of α -factor, a substantial amount of the total variation in system output was caused by differences in pathway subsystem output P (Table 1 and Supplementary Fig. 6a). At 1.25 nM α -factor, $\eta^2(P)$ accounted for 59% of the total output, but at 20 nM, $\eta^2(P)$ accounted for only 22% of the total (Table 1). This suggests that high levels of input might conceal pre-existing cell-to-cell differences in pathway capacity.

If expression capacity and pathway subsystem output were independent, then the larger variation in pathway subsystem output

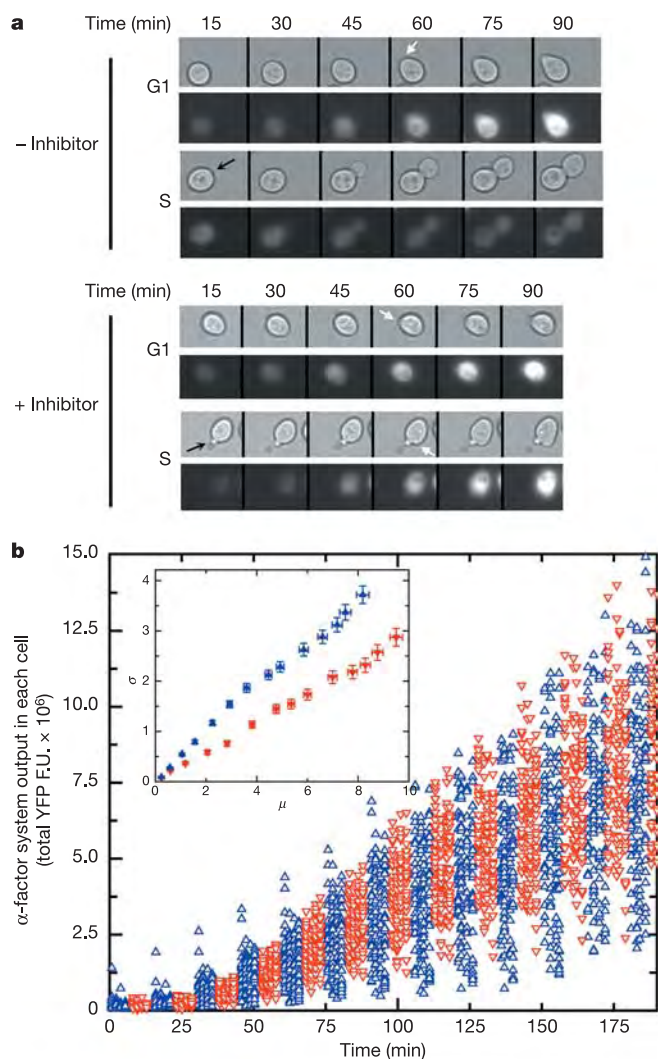


Figure 4 | The cell-cycle kinase Cdc28 causes a large part of the variation in α -factor system output. We treated TCY3154 cells (P_{PRM1} -YFP, P_{ACT1} -CFP, *cdc28-as2*) with 20 nM α -factor in the presence or absence of 10 μ M 1-NM-PP1 Cdc28-as2 inhibitor at time zero, collected YFP and CFP images, and quantified the fluorescent signals over time as in Fig. 2. **a**, Bright-field and YFP images of cells treated with 20 nM α -factor in the presence or absence of inhibitor, showing that Cdc28 inhibition allows cells in the S phase of the cell cycle to produce YFP in response to α -factor (black arrows mark the budding site). White arrows mark the first time point with a visible mating projection (shmoo tip) **b**, Alpha-factor system output of individual cells treated with 20 nM α -factor in the presence (red) or absence (blue) of 10 μ M inhibitor. For visual clarity, the red symbols have been shifted 8 min to the right. Inset shows the standard deviation (σ) of the population versus the median (μ) system output $\pm$ standard error on both axes (YFP F.U. $\times 10^6$).

Table 1 | α -factor concentration, Fus3 and Kss1 regulate cell-to-cell variation in pathway subsystem output

Strain	Promoters	α -factor (nM)	Total variation $\eta^2(y)$ ($\times 10^{-3}$)	Gene expression noise $\eta^2(\gamma_{YFP})$ ($\times 10^{-3}$)	Variation in pathway subsystem output $\eta^2(P_{YFP})$ ($\times 10^{-3}$)	Variation in G (+covariance) ($\times 10^{-3}$)	$\rho_{CFP,YFP}^*$
TCY3154 (WT)	P _{PRM1} -YFP versus P _{ACT1} -CFP	1.25	132 $\pm$ 12	5.62 $\pm$ 0.014 (4.26 $\pm$ 0.10)	78.1 $\pm$ 8.0 (59.2 $\pm$ 6.1)	48 $\pm$ 14 (36 $\pm$ 11)	0.725 $\pm$ 0.021
GPY3262 ($\Delta fus3$)	P _{PRM1} -YFP versus P _{ACT1} -CFP	1.25	125 $\pm$ 16	4.21 $\pm$ 0.14 (3.39 $\pm$ 0.11)	83.0 $\pm$ 11.0 (66.4 $\pm$ 9.2)	38 $\pm$ 19 (30 $\pm$ 15)	0.565 $\pm$ 0.049
GPY3263 ($\Delta kss1$)	P _{PRM1} -YFP versus P _{ACT1} -CFP	1.25	152 $\pm$ 17	2.76 $\pm$ 0.14 (1.82 $\pm$ 0.09)	54.6 $\pm$ 7.7 (35.8 $\pm$ 5.0)	95 $\pm$ 18 (62 $\pm$ 12)	0.774 $\pm$ 0.029
TCY3154 (WT)	P _{PRM1} -YFP versus P _{ACT1} -CFP	20	115 $\pm$ 12	2.01 $\pm$ 0.14 (1.74 $\pm$ 0.12)	24.8 $\pm$ 5.3 (21.6 $\pm$ 4.7)	88 $\pm$ 13 (77 $\pm$ 11)	0.867 $\pm$ 0.021
GPY3262 ($\Delta fus3$)	P _{PRM1} -YFP versus P _{ACT1} -CFP	20	128 $\pm$ 12	2.83 $\pm$ 0.14 (2.20 $\pm$ 0.11)	60.3 $\pm$ 7.1 (47.0 $\pm$ 5.5)	65 $\pm$ 14 (51 $\pm$ 11)	0.712 $\pm$ 0.031
GPY3263 ($\Delta kss1$)	P _{PRM1} -YFP versus P _{ACT1} -CFP	20	142 $\pm$ 24	2.40 $\pm$ 0.14 (1.693 $\pm$ 0.097)	36.9 $\pm$ 9.1 (26.0 $\pm$ 6.4)	103 $\pm$ 25 (72 $\pm$ 18)	0.827 $\pm$ 0.032

WT, wild type.

* Correlation coefficient between YFP and CFP.

Distribution of total cell-to-cell variation $\eta^2(y)$ among different sources. We treated strains TCY3154, GPY3262 and GPY3263 with the indicated α -factor concentration and 10 μ M 1-NM-PP1, and collected YFP and CFP images at 15-min intervals. The amount of variation due to the different sources of variation was calculated as explained in the main text and Supplementary Materials. Data correspond to the measurement 3 h after α -factor addition. The percentage of total variation $\eta^2(y)$ is given in parentheses. Error measurements are s.e.m. In addition to the shown errors, we associate a 10–15% systematic uncertainty with the reported numbers for the variation in pathway subsystem output and gene expression noise, due to the omission of higher-order terms in equation (2) (see Supplementary Materials).

$\eta^2(P)$ observed at low concentrations compared to high concentrations would sum with constant variation in expression capacity $\eta^2(G)$ and expression noise $\eta^2(\gamma)$ to result in a larger total variation $\eta^2(y)$ at low concentrations. Instead, $\eta^2(y)$ remained relatively constant with concentration (Table 1 and Supplementary Fig. S6). The fact that total variation does not increase significantly at low concentrations compared to high concentrations indicates that there is a negative correlation between G and L in equation (2), implying that cells with low expression capacity have a higher than expected pathway subsystem output and vice versa, and that cell-to-cell variation in pathways is regulated.

Fus3 and Kss1 regulate pathway power variation

Alpha-factor signal can be transmitted independently by two MAP kinases, Fus3 and Kss1 (ref. 17) (Fig. 1a). Fus3 and Kss1 both can phosphorylate the transcription factor Ste12, but they also phosphorylate distinct substrates^{25–27}. In addition, unphosphorylated Kss1 can bind and inhibit Ste12 (ref. 28). We reasoned that cell-to-cell differences in the relative levels of activated Fus3 and Kss1 might lead to different levels of active Ste12 and therefore to cell-to-cell

differences in pathway subsystem output. To test this idea, we derived strains from TCY3154 cells that lacked either Fus3 or Kss1, and stimulated the cells with high (20 nM) or low (1.25 nM) concentrations of α -factor. At high α -factor concentrations, the average responses of $\Delta fus3$ cells and $\Delta kss1$ cells were nearly the same as that of wild-type cells (within 20%). At low concentrations, the average response of $\Delta fus3$ cells was nearly the same as that of wild-type cells (within 5%). However, $\Delta kss1$ cells showed a stronger response than wild-type cells ($\sim$ 2-fold higher, data not shown), consistent with previous reports²⁹.

Relative to wild-type, $\Delta fus3$ cells had higher pathway variation at high α -factor concentrations and the same pathway variation at low concentrations. In contrast, relative to wild-type cells, $\Delta kss1$ cells showed the same pathway variation at high concentrations and lower pathway variation at low concentrations (Fig. 5b and Table 1). We performed western blots with an antibody that recognized the active forms of Fus3 and Kss1 (which are phosphorylated on two sites) to assess the relative activities of these protein kinases at high and low concentrations of α -factor. At high concentrations, we observed a larger phospho-Fus3 to phospho-Kss1 ratio than at low concentrations (data not shown), indicating that at high concentrations signal transmission depends more strongly on Fus3. Taken together, these results suggest that signal transmission is less variable when more dependent on Fus3 (at high α -factor concentrations or when Kss1 is absent) than when more dependent on Kss1 (at low concentrations or when Fus3 is absent).

Discussion

The approach described above is generally applicable to dissecting the sources of cell-to-cell variation for cellular processes for which the output can be measured with transcriptional reporters. Here we quantified the contributions of four sources of variation: cell-to-cell differences in pathway power or ability to transmit a signal (L), cell-to-cell differences in expression capacity or the ability to express proteins from genes (G), and noise in the operations of the pathway (λ) and the gene expression (γ) subsystems. We constructed an analytical framework to distinguish pathway activation from reporter gene expression, and used it to guide the design and interpretation of experiments. The framework relied on a subdivision of the process in general terms, rather than representation of the molecular mechanisms that underlie it. The use of such formalism has allowed us to model a non-steady-state process and to devise ways to measure pre-existing differences in the ability of cells to induce the pathway and express genes. In contrast to, for example, the rigorous treatment of mRNA and protein dynamics as coupled processes³⁰, we collapsed all

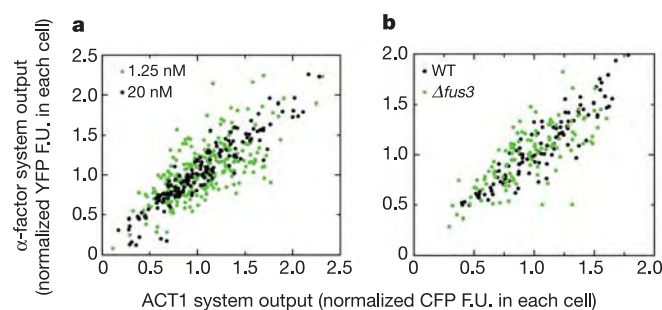


Figure 5 | Genetic control of cell-to-cell variation in pathway subsystem output $\eta^2(P)$. **a**, Alpha-factor regulates variation in pathway subsystem output. We treated TCY3154 cells with 20 nM (black) or 1.25 nM (green) α -factor and 10 μ M 1-NM-PP1. **b**, Fus3 reduces variation in pathway subsystem output. We treated TCY3154 (wild type, black) or GPY3262 ($\Delta fus3$, green) cells with 20 nM α -factor and 10 μ M 1-NM-PP1. Data correspond to the output of the α -factor system versus the ACT1 system 3 h after stimulation. YFP and CFP F.U. were normalized to the median of each population to allow the overlaying of data with different means. The increased variation in pathway subsystem output at low concentrations of α -factor in wild-type cells (**a**) and at high concentrations in $\Delta fus3$ cells (**b**) is manifested as a wider spread along the minor axis.

steps from reporter gene transcription to fluorescence output into a single composite process. This formalism can be modified as new experimental techniques allow the measurement of different processes, and as deeper molecular understanding leads us to further subdivide the causes of cell-to-cell variation.

We have shown that about half of the cell-to-cell variation is due to pre-existing differences in the cell-cycle position of individual cells at the time of pathway induction. This would be expected on the basis of previous work showing that, when yeast cells are about to commit to a new round of cell division, a complex between Cdc28 and Cln2 inhibits activation of the MAP kinase cascade in response to mating pheromone²². Our results extend this earlier work by suggesting that there are no relevant slow processes (such as transcription) mediating the Cdc28/Cln2-dependent inhibition of pathway activation and that the substrate(s) of Cdc28/Cln2 is relatively short-lived, as we observed the same reduction in variation whether we pre-incubated the cells with the Cdc28 inhibitor or added it simultaneously with α -factor (data not shown).

Another large component of the variation in system output is due to cell-to-cell differences in the capacity of cells to express proteins from genes (G), whereas little is due to noise in gene expression ($\eta^2(\gamma)$). Genetically identical cells had different G values, perhaps due to differences in the numbers of ribosomes or RNA polymerase II complexes, or the cellular energy level. Our findings, combined with recent studies in *E. coli*^{12,13}, suggest that variations in G are found in both eukaryotes and prokaryotes.

All processes that depend on levels of gene expression should be sensitive to cell-to-cell differences in expression capacity, as we showed for the process of cell volume changes. As *S. cerevisiae* cells initiate daughter formation when they reach a critical volume³¹, cells with high expression capacity should also reproduce more rapidly and account for a greater proportion of newly formed daughter cells in the population. However, we find the distribution of expression capacity in an exponentially growing population is stable over time (data not shown), suggesting that expression capacity is not strongly heritable, as reported for *E. coli*¹². Although expression capacity might change with cell age, our experiments show that age alone cannot account for high variation in expression capacity. In our exponentially growing populations most of the cells are young; approximately 60% are newborn daughters and half of the remainder (20% of the total) are cells that have given birth only once.

Notably, we found that variation in output of the pathway subsystem changed with α -factor input: at high concentrations, output variation was low, whereas at low concentrations, variation was high. We expected the noise component of this variation $\eta^2(\lambda)$ to behave like gene expression noise $\eta^2(\gamma)$, decreasing with increasing mean system output (Supplementary Fig. S4). However, the output variation of the pathway subsystem did not decrease with increasing mean system output over time, suggesting that pathway subsystem output variation is dominated by cell-to-cell differences in pathway power (L) rather than noise $\eta^2(\lambda)$.

Our results indicate that the amount of variation in the pathway subsystem output is regulated by the MAP kinases Fus3 and Kss1. The Fus3-dependent reduction in variation might be due to Fus3 autoregulatory negative feedback. For example, activated Fus3 induces the protein phosphatase Msg5 that dephosphorylates and inactivates Fus3 (ref. 32) but not Kss1 (ref. 33). Such feedback mechanisms would tend to equalize the levels of active Fus3 between cells. The Kss1-dependent increase in cell-to-cell variation might result from inputs to Kss1 from filamentation³⁴ and cell wall integrity³⁵ pathways. Thus, increasing the relative activity of Kss1 (compared to Fus3) may make the α -factor pathway more sensitive to variation in these other inputs. It may be advantageous for cells to regulate variation in cell-fate decisions. At high levels of α -factor, the decision to respond is clear; cells should respond as best as they can in order to mate; and pathway subsystem output depends predominantly on Fus3. At low levels, however, the decision to

respond might have to be dependent on factors in addition to the level of α -factor. Consequently, the pathway might rely more heavily on Kss1, which can integrate the α -factor pathway with other cellular information-processing pathways.

We imagine that cell-to-cell variation in general cellular capacities (such as expression capacity) creates circumstances that can distort the transmission of signals and provides selective pressure for the evolution of specialized, compensatory mechanisms that enable cells to generate less biased, more uniform responses. Such a compensatory regulatory mechanism might be responsible for the negative correlation we observed between expression capacity and pathway power. This correlation decreases the effect of differences in expression capacity on the α -factor system output. We predict that some of the 'feedback' and inhibitory genes that modulate pathway subsystem output¹⁷ might function in this compensatory mechanism.

We undertook this work as a step towards predicting the quantitative output of a cell-fate decision system in response to defined perturbations. Although many of the molecular components that comprise this system are known, the mechanisms that control its quantitative behaviour are not. Our experiments have defined two such mechanisms, and have begun to identify genes that affect their function. We hope that the combination of physiological experimentation enabled by new measurement tools and existing molecular and genetic methods will allow us to gain greater insight into the mechanisms that regulate quantitative variation. Understanding mechanisms that regulate global capability to express proteins from genes might have applicability to protein expression and the engineering of biological systems. Understanding the mechanisms that constrain variation in cell-fate decision systems might also enable new therapeutic interventions, for example to narrow the distribution of cellular responses to a pro-apoptotic anti-cancer drug.

METHODS

Nucleic acid and yeast manipulations were performed as previously described^{36,37}. Derivation of yeast strains (Supplementary Table S1) from YAS245-5C (ref. 38) and the protein methods used are detailed in the Supplementary Materials. Quantification of system output from single cells and measurements of cell-to-cell variation were performed using time-lapse fluorescent microscopy followed by data analysis using Cell-ID (A.G., A.C.-L., T.C., K. Benjamin & R.B., submitted manuscript).

Received 31 March; accepted 5 July 2005.

Published online 18 September 2005.

1. Sternberg, P. W. & Horvitz, H. R. Pattern formation during vulval development in *C. elegans*. *Cell* **44**, 761–772 (1986).
2. Priess, J. R. & Thomson, J. N. Cellular interactions in early *C. elegans* embryos. *Cell* **48**, 241–250 (1987).
3. Kimble, J. & Hirsh, D. The postembryonic cell lineages of the hermaphrodite and male gonads in *Caenorhabditis elegans*. *Dev. Biol.* **70**, 396–417 (1979).
4. Kimble, J. Alterations in cell lineage following laser ablation of cells in the somatic gonad of *Caenorhabditis elegans*. *Dev. Biol.* **87**, 286–300 (1981).
5. Karp, X. & Greenwald, I. Post-transcriptional regulation of the E/Daughterless ortholog HLH-2, negative feedback, and birth order bias during the AC/VU decision in *C. elegans*. *Genes Dev.* **17**, 3100–3111 (2003).
6. Doe, C. Q. & Goodman, C. S. Early events in insect neurogenesis. II. The role of cell interactions and cell lineage in the determination of neuronal precursor cells. *Dev. Biol.* **111**, 206–219 (1985).
7. Jan, Y. N. & Jan, L. Y. Maggot's hair and bug's eye: role of cell interactions and intrinsic factors in cell fate specification. *Neuron* **14**, 1–5 (1995).
8. Hoang, T. The origin of hematopoietic cell type diversity. *Oncogene* **23**, 7188–7198 (2004).
9. Delbrück, M. The burst size distribution in the growth of bacterial viruses (bacteriophages). *J. Bacteriol.* **50**, 131–135 (1945).
10. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000).
11. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
12. Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* **307**, 1962–1965 (2005).
13. Pedraza, J. M. & van Oudenaarden, A. Noise propagation in gene networks. *Science* **307**, 1965–1969 (2005).

14. Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science* **304**, 1811–1814 (2004).
15. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature Genet.* **31**, 69–73 (2002).
16. Blake, W. J., Kærn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422**, 633–637 (2003).
17. Dohlman, H. G. & Thorner, J. W. Regulation of G protein-initiated signal transduction in yeast: paradigms and principles. *Annu. Rev. Biochem.* **70**, 703–754 (2001).
18. Roberts, C. J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
19. Heiman, M. G. & Walter, P. Pmlp, a pheromone-regulated multispanning membrane protein, facilitates plasma membrane fusion during yeast mating. *J. Cell Biol.* **151**, 719–730 (2000).
20. Jenness, D. D., Burkholder, A. C. & Hartwell, L. H. Binding of alpha-factor pheromone to *Saccharomyces cerevisiae* cells: dissociation constant and number of binding sites. *Mol. Cell. Biol.* **6**, 318–320 (1986).
21. Bishop, A. C. *et al.* A chemical switch for inhibitor-sensitive alleles of any protein kinase. *Nature* **407**, 395–401 (2000).
22. Oehlen, L. J. & Cross, F. R. G1 cyclins CLN1 and CLN2 repress the mating factor response pathway at Start in the yeast cell cycle. *Genes Dev.* **8**, 1058–1070 (1994).
23. Oehlen, L. J. & Cross, F. R. Potential regulation of Ste20 function by the Cln1-Cdc28 and Cln2-Cdc28 cyclin-dependent protein kinases. *J. Biol. Chem.* **273**, 25089–25097 (1998).
24. Wassmann, K. & Ammerer, G. Overexpression of the G1-cyclin gene *CLN2* represses the mating pathway in *Saccharomyces cerevisiae* at the level of the MEKK Ste11. *J. Biol. Chem.* **272**, 13180–13188 (1997).
25. Breitkreutz, A., Boucher, L. & Tyers, M. MAPK specificity in the yeast pheromone response independent of transcriptional activation. *Curr. Biol.* **11**, 1266–1271 (2001).
26. Chou, S., Huang, L. & Liu, H. Fus3-regulated Tec1 degradation through SCFCdc4 determines MAPK signalling specificity during mating in yeast. *Cell* **119**, 981–990 (2004).
27. Bao, M. Z., Schwartz, M. A., Cantin, G. T., Yates, J. R. & Madhani, H. D. III Pheromone-dependent destruction of the Tec1 transcription factor is required for MAP kinase signalling specificity in yeast. *Cell* **119**, 991–1000 (2004).
28. Bardwell, L., Cook, J. G., Zhu-Shimoni, J. X., Voora, D. & Thorner, J. Differential regulation of transcription: repression by unactivated mitogen-activated protein kinase Kss1 requires the Dig1 and Dig2 proteins. *Proc. Natl Acad. Sci. USA* **95**, 15400–15405 (1998).
29. Sabbagh, W. Jr, Flatauer, L. J., Bardwell, A. J. & Bardwell, L. Specificity of MAP kinase signalling in yeast differentiation involves transient versus sustained MAPK activation. *Mol. Cell* **8**, 683–691 (2001).
30. Paulsson, J. Summing up the noise in gene networks. *Nature* **427**, 415–418 (2004).
31. Johnston, G. C., Pringle, J. R. & Hartwell, L. H. Coordination of growth with cell division in the yeast *Saccharomyces cerevisiae*. *Exp. Cell Res.* **105**, 79–98 (1977).
32. Doi, K. *et al.* MSG5, a novel protein phosphatase promotes adaptation to pheromone response in *S. cerevisiae*. *EMBO J.* **13**, 61–70 (1994).
33. Andersson, J., Simpson, D. M., Qi, M., Wang, Y. & Elion, E. A. Differential input by Ste5 scaffold and Msg5 phosphatase route a MAPK cascade to multiple outcomes. *EMBO J.* **23**, 2564–2576 (2004).
34. Madhani, H. D. & Fink, G. R. The riddle of MAP kinase signalling specificity. *Trends Genet.* **14**, 151–155 (1998).
35. Lee, B. N. & Elion, E. A. The MAPKKK Ste11 regulates vegetative growth through a kinase cascade of shared signalling components. *Proc. Natl Acad. Sci. USA* **96**, 12679–12684 (1999).
36. Ausubel, F. M. *et al.* (eds) *Current Protocols in Molecular Biology* (Wiley, New York, 2004).
37. Guthrie, C. & Fink, G. R. (eds) *Methods in Enzymology. Guide to Yeast Genetics and Molecular Biology* (Academic, San Diego, 1991).
38. Colman-Lerner, A., Chin, T. E. & Brent, R. Yeast Cbk1 and Mob2 activate daughter-specific genetic programs to induce asymmetric cell fates. *Cell* **107**, 739–750 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to D. Pincus for his help with the MAPK phosphorylation measurements and to L. Lok, K. Benjamin, I. Burbulis and R. Yu for discussions and comments on the manuscript. Work was under the 'Alpha Project' at the Center for Genomic Experimentation and Computation, an NIH Center of Excellence in Genomic Science. The Alpha Project is supported by a grant from the National Human Genome Research Institute to R.B.

Author Contributions A.C.-L. and A.G. conceived the framework, developed the experimental methods, performed most of the experiments and analysed the results. R.B. provided input regarding problem choice, experimentation and interpretation. A.C.-L., A.G. and R.B. wrote the paper and stand as guarantors of its findings. T.C. made most of the plasmid and yeast strains. E.S. made some plasmids and yeast strains, and some of the measurements in Fig. 2. C.G.P. made the observation that Fus3 regulates pathway variation and collaborated in interpreting its biological implications. O.R. and A.C.-L. made the observation that activated Fus3/Kss1 ratios are α -factor-dependent. D.E. suggested analysis of the stochastic fluctuations in the system and helped with describing the framework.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.C.-L. (colman-lerner@molsci.org), A.G. (gordon@molsci.org) and R.B. (brent@molsci.org).

Lost and found dark matter in elliptical galaxies

A. Dekel^{1,2,3,4}, F. Stoehr², G. A. Mamon^{2,3}, T. J. Cox⁶, G. S. Novak⁵ & J. R. Primack⁴

There is strong evidence that the mass of the Universe is dominated by dark matter, which exerts gravitational attraction but whose exact nature is unknown. In particular, all galaxies are believed to be embedded in massive haloes of dark matter^{1,2}. This view has recently been challenged by the observation of surprisingly low random stellar velocities in the outskirts of ordinary elliptical galaxies, which has been interpreted as indicating a lack of dark matter^{3,4}. Here we show that the low velocities are in fact compatible with galaxy formation in dark-matter haloes. Using numerical simulations of disk-galaxy mergers^{5,6}, we find that the stellar orbits in the outer regions of the resulting ellipticals are very elongated. These stars were torn by tidal forces from their original galaxies during the first close passage and put on outgoing trajectories. The elongated orbits, combined with the steeply falling density profile of the observed tracers, explain the observed low velocities even in the presence of large amounts of dark matter. Projection effects when viewing a triaxial elliptical can lead to even lower observed velocities along certain lines of sight.

The common spiral galaxies are known to reside in extended dark-matter haloes. The rotational speeds of their gas disks do not decline outside the visible disk¹, unlike the expectation from keplerian velocities at a radius r about a mass M , $V^2 = GM/r$ (where G is Newton's constant). Thus, the dark-matter mass within r is growing roughly as $M(r) \propto r$ and it dominates the gravitational potential beyond a certain radius. An extrapolation based on the typical halo density profile⁷ found in simulations of the standard Λ CDM cosmology predicts an outer 'virial' radius R_{vir} that is 50–100 times larger than the characteristic stellar radius, enclosing 10–20 times more dark matter than luminous matter. The conventional wisdom is that the potential wells created by the dark matter are crucial for seeding the formation of galaxies^{2,8,9}.

The standard hypothesis is that elliptical galaxies originate from mergers of disks¹⁰ and should therefore be embedded in similar dark-matter haloes. There is evidence for dark matter in giant ellipticals, from X-rays¹¹ and gravitational lensing¹². However, ordinary ellipticals lack obvious velocity tracers at the large projected radii r_p where the dark matter is expected to be important. This is typically beyond R_{eff} (ref. 13), the 'effective' radius encompassing half the light, while measurements of the projected velocity dispersion σ_p of the stellar light are limited to $r_p < 2R_{\text{eff}}$.

The strong [O III] emission line at 5,007 Å from planetary nebulae—hot shells of gas expelled from dying stars of $(1\text{--}3)M_{\odot}$ —provides a unique tool for extracting $\sigma_p(r_p)$ beyond R_{eff} . The σ_p of planetary nebulae in the normal ellipticals NGC 821, 3379 and 4494 (ref. 4) and in NGC 4697 (ref. 3) were found typically to drop by a factor ≈ 1.6 between $r_p = R_{\text{eff}}$ and $3R_{\text{eff}}$. Kinematic modelling by the observers⁴ yielded low mass-to-light ratios, for example $M/L \approx 7$ at $5R_{\text{eff}}$ for NGC 3379, consistent with a "naked" stellar population. They interpreted this as "little if any dark matter in these galaxies' haloes". While noticing that increasing velocity anisotropies could in

principle produce declining σ_p , they ruled out such "pathological" orbit structure. Similar conclusions were obtained later from other ellipticals¹⁴. The apparent challenge to theory has already triggered radical explanations¹⁵. However, the earlier analysis⁴ might have missed alternative solutions because it was limited to specific density-profile shapes in stationary spherical systems, the halo planetary nebulae were identified with the central stellar population, and their maximum-likelihood method may suffer from an incomplete orbit library or questionable convergence properties¹⁶.

For given density profiles, a lower σ_p can result from more radial velocities. The dynamics implies a lower three-dimensional velocity dispersion σ because the pressure needed for balancing gravity is provided by a radial σ_r that corresponds to a lower σ . The projection introduces a further decrease in σ_p . This can be illustrated by toy models made purely of circular or radial orbits, with the same constant speed and random orientations. If the stellar-density profile is steep enough, σ_p is dominated by the tangential contribution near the equatorial plane perpendicular to the line of sight, which is high for circular orbits and low for radial orbits.

The three-dimensional profiles of any component of a spherical gravitating system in equilibrium obey the Jeans equation¹⁷:

$$V^2(r) = [\alpha(r) + \gamma(r) - 2\beta(r)]\sigma_r^2(r) \quad (1)$$

a manifestation of local hydrostatic balance between the inward pull of gravity (left) and the outward push of pressure (right). Here $V^2(r) = GM(r)/r$ is the squared circular velocity. The stellar density profile $\nu(r)$ enters via $\alpha \equiv -d\ln\nu/d\ln r$. Its velocity dispersion consists of radial and tangential components, $\sigma^2 = \sigma_r^2 + 2\sigma_\theta^2$; we define $\gamma \equiv -d\ln\sigma_r^2/d\ln r$. The velocity anisotropy is $\beta \equiv 1 - \sigma_\theta^2/\sigma_r^2$, with $\beta = -\infty, 0, 1$ for circular, isotropic and radial orbits respectively. The projection can be performed analytically when β , α and γ are constant with r (power-law profiles, $V^2 = V_0^2(r/R_{\text{eff}})^{-\gamma}$):

$$\sigma_p^2(r_p) = A(\alpha, \gamma) \left(\frac{(\alpha + \gamma) - (\alpha + \gamma - 1)\beta}{(\alpha + \gamma) - 2\beta} \right) V_0^2 \left(\frac{r_p}{R_{\text{eff}}} \right)^{-\gamma} \quad (2)$$

(see Methods for definition of A). We note that σ_p is a decreasing function of β and of α (for $\alpha + \gamma > 3$ and $\beta > 0$). Local fits to the de-projection of the standard de Vaucouleurs¹⁸ surface-brightness profile of ellipticals give $\alpha \approx 3.13\text{--}3.37$ at $2\text{--}3R_{\text{eff}}$ (ref. 19). Our fits to $\sigma_p^2(r_p)$ in the observed ellipticals beyond R_{eff} yield $\gamma \approx 0.8 \pm 0.2$. These give $A(\alpha, \gamma) \approx 0.2$ and σ_p drops by a factor of ~ 1.5 between $\beta = 0$ and 1. We learn that one could match the low σ_p at large radii either by a low V_0 or by a high β there, and that a high α helps.

In a more realistic model, we assume a Sérsic stellar density profile¹⁹ (equivalent to that of de Vaucouleurs for Sérsic index $m = 4$), a standard dark-matter density profile⁷ with a typical concentration (~ 10) (ref. 20), and a virial stellar mass fraction of ~ 0.06 . With $\beta = 0$ we recover the discrepancy, but $m \approx 4$ and $\beta(r > R_{\text{eff}}) \approx 0.5$ (independent of its behaviour well inside R_{eff}) yield a

¹Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel. ²Institut d'Astrophysique, 98bis Boulevard Arago, Paris 75014, France. ³Observatoire de Paris, F-92195 Meudon, France. ⁴Department of Physics, ⁵UCO/Lick Observatories, University of California, Santa Cruz, California 95064, USA. ⁶Center for Astrophysics, Harvard University, 60 Garden Street, Cambridge, Massachusetts 02138, USA.

$\sim 1\sigma$ agreement with the observed σ_p . A good fit is obtained with either $m \approx 4$ and $\beta \approx 0.75$, or with $m \approx 2.3$ (for which $\alpha \approx 3.5$ near $2.5R_{\text{eff}}$) and $\beta \approx 0.5$. The required β is higher than the $\beta \leq 0.2$ predicted for dark matter in typical haloes¹³, so the planetary nebulae must not trace the dark-matter kinematics.

Assuming that ellipticals form by mergers, and that major mergers of disks can reveal generic features of mergers in the Λ CDM cosmology, we appeal to a suite of simulations of such events^{5,6}. Two spirals are put on a parabolic orbit, each consisting of stellar and gaseous disks and a bulge, all embedded in a Λ CDM halo, constructed to match a range of typical disk galaxies. The gravitational and hydrodynamical evolution is followed using an SPH code²¹, including gas cooling, star formation and supernova feedback (Methods).

Figure 1 shows the stacked three-dimensional profiles of the merger remnants. The dark-matter density profile is slightly flatter than isothermal, $\rho \propto r^{-2}$, similar to simulated Λ CDM haloes after they have responded to gas dissipation²². The robust stellar density falls off more steeply, $\rho \propto r^{-3.2}$, as in ellipticals obeying the de Vaucouleurs profile near $2\text{--}3R_{\text{eff}}$ and with $R_{\text{eff}} \approx 0.015R_{\text{vir}}$. For the “young” stars it is somewhat steeper: $\rho \propto r^{-3.5}$. The total-to-stellar mass ratio rises from ≈ 2 at $3R_{\text{eff}}$ to ≈ 14 at R_{vir} , corresponding at $5R_{\text{eff}}$ to $M/L \approx 15$ (compared to the earlier⁴ $M/L \approx 7$). The three-dimensional σ profiles of the dark matter and stars have similar slopes (as in equation (2)), roughly $\sigma \propto r^{-0.2}$.

Our main finding is the high β of the stellar halo velocities. While the dark-matter velocities are almost isotropic ($\beta \approx 0.1$), the typical stellar β grows from small values at $r < R_{\text{eff}}$ (sometimes negative, reflecting a small disk that is irrelevant beyond R_{eff}) to $\beta \approx 0.5$ at $r > R_{\text{eff}}$. In one case $\beta > 0.75$, but in another it remains < 0.2 . Given $V(r)$ in the Jeans equation, the higher α is compensated by a higher β and a lower σ .

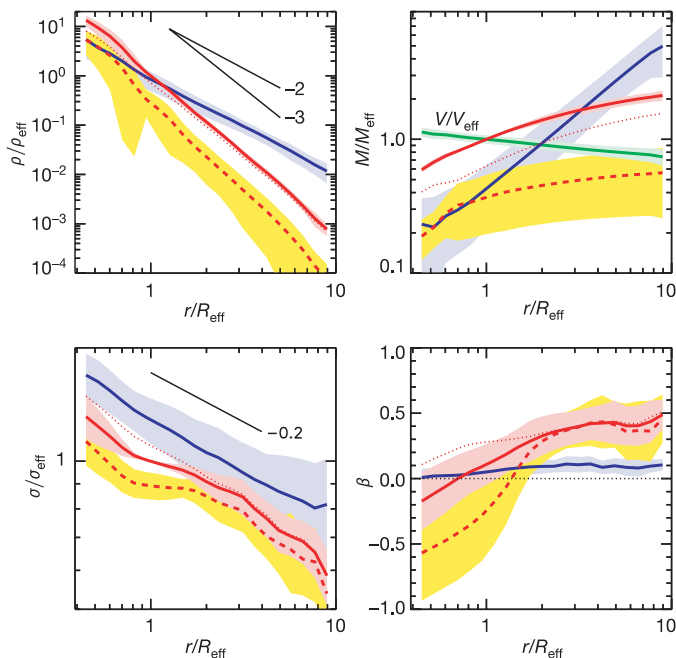


Figure 1 | Three-dimensional profiles of the simulated merger remnants. Ten galaxies at two different times after the merger (typically 0.8 and 1.3 Gyr) are stacked. Shown are the profiles for the dark matter (blue) and the stars (red), divided into the “old” stars from the progenitors (dotted) and the “young” stars formed during the merger (dashed). The scaling is such that the curves for the stars (“all”, solid red) are matched at R_{eff} . The shaded areas mark 1σ scatter. The panels refer to density ρ , mass M and circular velocity V , velocity dispersion σ and anisotropy β , with subscript ‘eff’ referring to R_{eff} .

The simulations demonstrate that the stellar halo originates from tidal processes during the first pericentre passage. Some of the halo stars are associated with the two cores and the tidal bridge between them; they pass near the centre at the coalescence before flying outward on radial orbits. Other halo stars first flow out in extended tidal tails and later fall back on radial orbits (Supplementary Information). Indeed, we find that β is correlated with the strength of the tidal interaction; it is higher for more head-on collisions and when the spins are aligned with the orbit.

The systems are ‘observed’ from three orthogonal directions and stacked together, providing a robust average profile and the scatter about it. The data are scaled similarly (see Methods). Figure 2 shows the simulated surface-density profile and those of NGC 821 (ref. 23), NGC 3379 (ref. 24) and NGC 4697 (ref. 25), all fitted by $\Sigma \propto r_p^{-2.3}$ in $1\text{--}5R_{\text{eff}}$ (as in the de Vaucouleurs profile at $\sim 2R_{\text{eff}}$). The simulated projected axial ratios near R_{eff} range from 1:1 to 1:2, and the ellipticity is supported by an anisotropic, triaxial velocity dispersion rather than by rotation, similar to ellipticals (Supplementary Information). The distribution of global properties of the remnants, such as luminosity, radius and velocity dispersion, is consistent with the ‘fundamental plane’ of ellipticals (Supplementary Information). Thus, the merger remnants seem to resemble typical ellipticals near $\sim R_{\text{eff}}$ in every relevant respect.

The velocity dispersions in Fig. 2 illustrate our main point. While the dark-matter σ_p indeed lies above the outer observed points, the stellar σ_p , $\sim 30\%$ lower, provides a good fit everywhere in the range

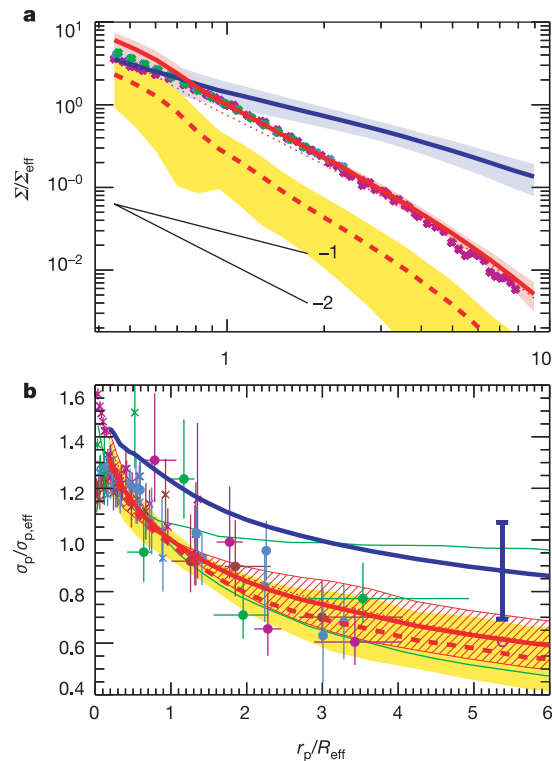


Figure 2 | Projected profiles: simulated galaxies versus observations.

a, Surface density. **b**, Velocity dispersion. The merger remnants are viewed from three orthogonal directions and the 60 profiles are stacked such that the curves for “all” the stars match at R_{eff} . Colours and line types are as in Fig. 1. The < 3 -Gyr “young” stars may mimic the observed planetary nebulae. The 1σ scatter is marked by a hashed area (“all”), a shaded area (“young”) or a thick bar (dark matter). The galaxies are marked green (NGC 821), violet (NGC 3379), brown (NGC 4494) and blue (NGC 4697) with 1σ errors; planetary nebulae (circles) and stars (crosses). The surface densities shown for three galaxies almost coincide with the simulated profile. Green lines refer to earlier models⁴ with (upper) and without (lower) dark matter.

0.5–4 R_{eff} . The slope of the simulated $\sigma_p^2(r_p)$ in the range 0.5–6 R_{eff} is $\gamma = 0.53 \pm 0.16$ for “all” and $\gamma = 0.61 \pm 0.22$ for the “young” stars, both consistent with the $\gamma = 0.59 \pm 0.13$ observed for planetary nebulae. No adjustment of model parameters is involved—simply stacking a sample of merger remnants.

The σ_p of “young” stars is lower by $\sim 9\%$ at $3R_{\text{eff}}$ (owing to their larger α). Stellar theory indicates that these objects, < 3 Gyr old, may represent the observed planetary nebulae. Emerging from 1.4–2.5 $M_{\odot}$ stars²⁶, the nebulae are expected to be much more luminous than those of the older, less massive stars, which fall below the detection limits^{3,4} (Methods). However, the radial orbits are a generic result independent of the degree of gas dissipation during the merger: our mergers with initial gas-to-baryon ratio ranging from 0 to 70% show negligible differences in β beyond R_{eff} . Dissipation results in a more centrally concentrated stellar distribution, associated with $\lesssim 10\%$ reduction in outer σ_p (Supplementary Information). We also find that the radial orbits and low σ_p emerge from major and minor mergers alike, independent of the progenitor mass ratio (Supplementary Information), and that the presence of a $\sim 22\%$ bulge does not make a significant difference. The tidal origin of the stellar halo explains this robustness to many merger characteristics. Furthermore, one simulation continued till 3.5 Gyr after the merger with no sign of evolution in $\beta(r)$ and $\sigma_p(r)$ beyond $\sim R_{\text{eff}}$ (Supplementary Information).

The $\pm 20\%$ scatter in σ_p is partly due to the angular momentum of the merger orbit and the relative spin inclinations, but also due to the line of sight relative to the principal axes of the triaxial system (or rotation axis, A. Burkert, R. Kudritzki and R. Mendez, manuscript in preparation). When viewed ‘face-on’, some of the remnants show σ_p values lower than observed, whereas other extreme ‘edge-on’ cases show σ_p values almost as high as the dark matter (Supplementary Information).

The simulated line-of-sight velocity distribution is also consistent with the data beyond the second moment. The deviations from gaussian are small, with the fourth moment $h_4 = 0.03 \pm 0.05$ for the central stars (Supplementary Information). At $r \gtrsim R_{\text{eff}}$, radial, prograde mergers produce negligible h_4 values, as in the planetary nebulae of NGC 3379 (ref. 4) (or small positive values as in NGC 5128; ref. 27), whereas more circular and retrograde mergers, or gas-rich mergers, can produce negative h_4 as in NGC 4697 (ref. 3; Supplementary Information).

We conclude that the planetary-nebulae data are consistent with the simple picture where normal ellipticals also reside in massive dark-matter haloes. The low σ_p is primarily due to the radial orbits of the halo stars, being tidally ejected from the inner regions during mergers independently of dissipation and mass ratio. This generic origin of the radial orbits indicates that the results based on our sample of simulations are representative of a broader range of merger types expected in the Λ CDM cosmology, and argues that the low observed planetary-nebula velocities are a natural outcome of this standard model. This should be confirmed by cosmological simulations (such as low-resolution results^{28,29}).

The range of merger properties leads to a variety of β and σ_p profiles, and the triaxiality adds directional variations, allowing extreme σ_p values smaller and larger than the planetary-nebula data. The possible association of the planetary nebulae with the younger stars, whose density profile is slightly steeper, may help reducing their σ_p a bit further. Other tracers, involving old stars, are expected to show a somewhat higher σ_p . This is especially true for globular clusters¹⁴ (G. Bergond, S. E. Zepf, A. J. Romanowsky, R. M. Sharples & K. L. Rhode, manuscript in preparation), given their flatter density profile³⁰ and presumably lower anisotropy (due to tidal disruption in radial orbits). A somewhat higher σ_p may also be expected in elliptical–elliptical mergers, common especially in groups (as observed¹⁴), where the collision orbits may be circularized by two-body relaxation and dynamical friction.

We recall⁴ that a nearly isotropic “naked” stellar system also

provides a fit to the $\sigma_p(r_p)$ data inside a few R_{eff} , appealing to a low V_0 rather than a high β in equation (2). Our simulations provide another solution, which does include a standard dark-matter halo. The dark-matter model predicts that σ_p flattens toward $\sim 10R_{\text{eff}}$ (as in NGC 5128; ref. 27), where the “naked” model predicts a continuing decline beyond $3R_{\text{eff}}$ and very low σ_p for other tracers as well. Whereas the “naked” model violates much of what we know about galaxy formation and cosmology, the dark-matter model, with the radial stellar orbits, seems to be a straightforward outcome of the self-consistent picture of structure formation in the Universe.

METHODS

Equation 2 derivation. The coefficient in equation (2) is a weak function of α and γ :

$$A(\alpha, \gamma) = \frac{1}{(\alpha + \gamma)} \frac{\Gamma[(\alpha + \gamma - 1)/2]}{\Gamma[(\alpha + \gamma)/2]} \frac{\Gamma[(\alpha/2)]}{\Gamma[(\alpha - 1)/2]} \quad (3)$$

For instance, $A(3.5, 1.0) \approx 0.18$, while $A(3.0, 0.4) \approx 0.26$.

Merger simulations. The simulations^{5,6} represent some of the major collisions that probably occurred during the hierarchical structure formation according to the Λ CDM cosmology. The evolution is followed using the entropy-conserving, gravitating, smoothed particle hydrodynamics (SPH) code GADGET²¹. Gas cooling, star formation and supernova feedback are treated using recipes that were calibrated to match observed star-formation rates.

The progenitor disk galaxies mimic typical big spirals: one type (G) representing today’s Sb galaxies and another type containing more gas as in Sbc-Sc galaxies and at high redshift. The sample consists of four G mergers, with dark-matter masses $1.2 \times 10^{12} M_{\odot}$ (except one $5 \times 10^{11} M_{\odot}$), and five Sbc merger plus one Sc merger with $8 \times 10^{11} M_{\odot}$ haloes. The baryonic fraction is $\sim 5\%$ of the dark-matter halo mass in the G cases, and $\sim 13\%$ in the Sbc-Sc cases. The fraction of baryons in gas is $\sim 20\%$ in G, 52% in Sbc, and 70% in Sc. The particle mass is $< 10^6 M_{\odot}$ for gas and stars and $\lesssim 10^7 M_{\odot}$ for dark-matter. The smoothing length h is 100 and 400 pc respectively, with the force becoming newtonian at $\geq 2.3 h$.

Two identical galaxies are set on parabolic orbits and merge because of dynamical friction due to their massive haloes. Our sample consists of several different orbits and orientations, including prograde and retrograde configurations in which the galaxy spins are aligned or antialigned with the orbital angular momentum. The merger results in two successive starbursts, one after the first close approach, and the other after the second, final coalescence (Supplementary Information). The starbursts occur 1–2 Gyr after the beginning of the simulation, and the remnant is ‘observed’ ~ 1 Gyr later. The amount of stars formed during the merger is roughly proportional to the initial gas fraction, and is not very sensitive to the orbit or orientation. The instantaneous rate is $10\text{--}100 M_{\odot} \text{ yr}^{-1}$. The young stars formed during the merger constitute $\sim 30\%$ of the total stars; typically 20% in the G galaxies and 40% in Sbc galaxies. The remnant galaxies resemble normal elliptical galaxies, as demonstrated above (Supplementary Information).

Scaling the data. The observed galaxies are presented together using the R_{eff} of each surface-brightness profile^{4,24,25}. We note that R_{eff} for NGC 3379 (ref. 24, 25) is $\approx 50\%$ larger than that quoted⁴. In Fig. 2, an open circle marks the last point had we used the smaller⁴ R_{eff} , not making a qualitative difference. The amplitudes of Σ and σ_p are scaled by least-squares fits of the stellar data at $r > 0.2R_{\text{eff}}$ to the stacked simulated profile as a reference. Replacing this reference by a different function of a similar general shape yields similar results. Using only the stars at larger radii (up to $r > R_{\text{eff}}$), or using the planetary nebulae alone, yield σ_p adjusting factors that differ only by a few per cent.

The R_{eff} of NGCs 821, 3379, 4494 and 4697 match the mean simulation value after multiplication by 0.667, 1.57, 1.00 and 1.13, indicating that the simulated and observed galaxies are of comparable sizes. The σ_p were adjusted by factors 1.00, 1.19, 1.21 and 1.11 for best fit. Being comparable to the radius scaling factors indicates that the observed and simulated galaxies have a similar velocity structure. The mean and 1σ scatter in the simulated remnants are $R_{\text{eff}} = 4.05 \pm 1.04 \text{ kpc}$ and $\sigma_p(R_{\text{eff}}) = 154 \pm 33 \text{ km s}^{-1}$.

Age of planetary nebulae. The [O III] luminosity of a planetary nebula with mass $< 2.5 M_{\odot}$ is strongly increasing with the parent stellar mass (figures 10 and 14 in ref. 26), hence sharply decreasing with age. A limiting magnitude M_{5007} then corresponds to a maximum stellar age t . For the complete sample of 328 planetary nebulae in NGC 4697³ it is $M_{5007} \approx -2.6$, namely $t_{M01} < 3$ Gyr (figures 18, 19, 26 in ref. 26). With only ≈ 100 planetary nebulae per galaxy⁴, the magnitude limit is brighter (by -0.8 magnitudes based on telescope gathering areas), so the stars are even younger. Based on theoretical planetary

nebulae luminosity functions (figures 18, 26 in ref. 26), if the population is typically older than 1 Gyr, then $t_{R03} < 2$ Gyr. We therefore adopt $t < 3$ Gyr as a limit for most of the observed planetary nebulae in the four galaxies. This indicates an association with the “young” simulated stars, and that the mergers of gaseous disks are relevant to those ellipticals showing planetary nebulae. A caveat is the apparent relative invariance of the planetary-nebula luminosity function between galaxies, seemingly independent of signs for a recent major merger. When there are no such signs, the observed planetary nebulae may be the signature of recent minor mergers, which are expected to produce similar effects.

Received 27 January; accepted 22 June 2005.

- Sofue, Y. & Rubin, V. Rotation curves of spiral galaxies. *Annu. Rev. Astron. Astrophys.* **39**, 137–174 (2001).
- White, S. D. M. & Rees, M. J. Core condensation in heavy halos—A two-stage theory for galaxy formation and clustering. *Mon. Not. R. Astron. Soc.* **183**, 341–358 (1978).
- Méndez, R. H. *et al.* Detection, photometry, and slitless radial velocities of 535 planetary nebulae in the flattened elliptical galaxy NGC 4697. *Astrophys. J.* **563**, 135–150 (2001).
- Romanowsky, A. J. *et al.* A dearth of dark matter in ordinary elliptical galaxies. *Science* **301**, 1696–1698 (2003).
- Cox, T. J. *Simulations of Galaxy Mergers: Star Formation and Feedback*. PhD thesis, UC Santa Cruz (2004); <http://physics.ucsc.edu/~tj/work/thesis/>.
- Cox, T. J., Jonsson, P., Primack, J. R. & Somerville, R. S. The effects of feedback in simulations of disk galaxy major mergers. Preprint at (<http://arXiv.org/astro-ph/0503201>) (2005).
- Navarro, J. F., Frenk, C. S. & White, S. D. M. A Universal Density Profile from Hierarchical Clustering. *Astrophys. J.* **490**, 493–510 (1997).
- Blumenthal, G. R., Faber, S. M., Primack, J. R. & Rees, M. J. Formation of galaxies and large-scale structure with cold dark matter. *Nature* **311**, 517–525 (1984).
- Dekel, A. & Silk, J. The origin of dwarf galaxies, cold dark matter, and biased galaxy formation. *Astrophys. J.* **303**, 39–55 (1986).
- Fall, S. M. Dissipation, merging and the rotation of galaxies. *Nature* **281**, 200–202 (1979).
- Mathews, W. G. & Brighenti, F. Hot gas in and around elliptical galaxies. *Annu. Rev. Astron. Astrophys.* **41**, 191–239 (2003).
- Keeton, C. R. Cold dark matter and strong gravitational lensing: concord or conflict? *Astrophys. J.* **561**, 46–60 (2001).
- Mamon, G. A. & Lokas, E. L. Dark matter in elliptical galaxies: II. Estimating the mass within the virial radius. *Mon. Not. R. Astron. Soc.* (in the press); preprint at (<http://arXiv.org/astro-ph/0405491>) (2005).
- Napolitano, N. R. *et al.* Mass-to-light ratio gradients in early-type galaxy haloes. *Mon. Not. R. Astron. Soc.* **357**, 691–706 (2005).
- Milgrom, M. & Sanders, R. H. Modified newtonian dynamics and the “dearth of dark matter in ordinary elliptical galaxies”. *Astrophys. J. Lett.* **599**, L25–L28 (2003).
- Valluri, M., Merritt, D. & Emsellem, E. Difficulties with recovering the masses of supermassive black holes from stellar kinematical data. *Astrophys. J.* **602**, 66–92 (2004).
- Binney, J. & Tremaine, S. *Galactic Dynamics* Ch. 4.2.d, eq. 4–55 (Princeton Univ. Press, Princeton, NJ, 1987).
- de Vaucouleurs, G. Recherches sur les Nebuleuses Extragalactiques. *Ann. Astrophys.* **11**, 247–287 (1948).
- Lima Neto, G. B., Gerbal, D. & Marquez, I. The specific entropy of elliptical galaxies: an explanation for profile-shape distance indicators. *Mon. Not. R. Astron. Soc.* **309**, 481–495 (1999).
- Bullock, J. S. *et al.* Profiles of dark haloes: evolution, scatter and environment. *Mon. Not. R. Astron. Soc.* **321**, 559–575 (2001).
- Springel, V., Yoshida, N. & White, S. D. M. GADGET: a code for collisionless and gas dynamical cosmological simulations. *New Astron.* **6**, 79–117 (2001).
- Gnedin, O. Y., Kravtsov, A. V., Klypin, A. A. & Nagai, D. Response of dark matter halos to condensation of baryons: cosmological simulations and improved adiabatic contraction model. *Astrophys. J.* **616**, 16–26 (2004).
- Goudfrooij, P. *et al.* Interstellar matter in Shapley-Ames elliptical galaxies. I. Multicolour CCD surface photometry. *Astron. Astrophys. Suppl.* **104**, 179–231 (1994).
- de Vaucouleurs, G. & Capaccioli, M. Luminosity distribution in galaxies. I—The elliptical galaxy NGC 3379 as a luminosity distribution standard. *Astrophys. J. Suppl.* **40**, 699–731 (1979).
- Peletier, R. F., Davies, R. L., Illingworth, G. D., Davis, L. E. & Cawson, M. CCD surface photometry of galaxies with dynamical data. II—UBR photometry of 39 elliptical galaxies. *Astron. J.* **100**, 1091–1142 (1990).
- Marigo, P., Girardi, L., Weiss, A., Groenewegen, M. A. T. & Chiosi, C. Evolution of planetary nebulae. II. Population effects on the bright cut-off of the PNLF. *Astron. Astrophys.* **423**, 995–1015 (2004).
- Peng, E. W., Ford, H. C. & Freeman, K. C. The planetary nebula system and dynamics in the outer halo of NGC 5128. *Astrophys. J.* **602**, 685–704 (2004).
- Sáiz, A., Domínguez-Tenreiro, R. & Serna, A. Elliptical galaxies at $z = 0$ from self-consistent hydrodynamical simulations: comparison with Sloan Digital Sky Survey structural and kinematical data. *Astrophys. J. Lett.* **601**, L131–L134 (2004).
- Abadi, M. G., Navarro, J. F. & Steinmetz, M. Stars beyond galaxies: The origin of extended luminous halos around galaxies. Preprint at (<http://arXiv.org/astro-ph/0506659>) (2005).
- Côté, P. *et al.* Dynamics of the globular cluster system associated with M87 (NGC 4486). II. Analysis. *Astrophys. J.* **559**, 828–850 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge discussions with M. Beasley, A. Burkert, K. Gebhardt, J. Navarro, A. Romanowsky and his group, and assistance from M. Covington. This research has been supported by the Israel Science Foundation and by NASA and NSF at UCSC. The simulations were run at NERSC. A.D. acknowledges a Miller Professorship at UC Berkeley, support from UCO/Lick Observatory, and a Blaise Pascal International Chair in Paris.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.D. (dekel@phys.huji.ac.il).

Isotope-induced partial localization of core electrons in the homonuclear molecule N₂

Daniel Rolles¹, Markus Braune¹, Slobodan Cvejanović^{1†}, Oliver Gefñner^{1†}, Rainer Hentges¹, Sanja Korica¹, Burkhard Langer^{2†}, Toralf Lischke^{1†}, Georg Prümper^{1†}, Axel Reinköster¹, Jens Viehhaus¹, Björn Zimmermann^{3†}, Vincent McKoy³ & Uwe Becker¹

Because of inversion symmetry and particle exchange, all constituents of homonuclear diatomic molecules are in a quantum mechanically non-local coherent state; this includes the nuclei and deep-lying core electrons. Hence, the molecular photoemission can be regarded as a natural double-slit experiment¹: coherent electron emission originates from two identical sites, and should give rise to characteristic interference patterns². However, the quantum coherence is obscured if the two possible symmetry states of the electronic wavefunction ('*gerade*' and '*ungerade*') are degenerate; the sum of the two exactly resembles the distinguishable, incoherent emission from two localized core sites. Here we observe the coherence of core electrons in N₂ through a direct measurement of the interference exhibited in their emission. We also explore the gradual transition to a symmetry-broken system of localized electrons by comparing different isotope-substituted species—a phenomenon analogous to the acquisition of partial 'which-way' information in macroscopic double-slit experiments³.

With respect to molecular inversion symmetry, the electronic wavefunctions of homonuclear diatomic molecules can be described as symmetry-adapted linear combinations of the corresponding atomic wavefunctions *a* and *b*, a situation actually realized by imposition of a fixed phase between them. For the core electrons, these symmetry-adapted wavefunctions Ψ , both *gerade* (*g*) and *ungerade* (*u*), can be written as:

$$\Psi_g = 1/\sqrt{2(1+S)} \times [\Psi_a(r) + \Psi_b(r)] \text{ and}$$

$$\Psi_u = 1/\sqrt{2(1-S)} \times [\Psi_a(r) - \Psi_b(r)]$$

respectively, where the phases of the two orbitals in Ψ_u differ by π and the overlap-integral is given by $S = \int \Psi_a(r) \Psi_b(r) dr$. The corresponding molecular orbitals for K-shell electrons are designated as $1\sigma_g$ and $1\sigma_u$.

The experimental fingerprint of the coherence of photoelectron emission from the $1\sigma_g$ and $1\sigma_u$ states in molecular nitrogen is the angular distributions in the molecular frame. These are predicted to exhibit characteristic differences between the two symmetry states, particularly in their nodal structure, which reflect the angular-momentum-dependent partial wave composition of the photoelectron wavefunction. In the non-local coherent case, this composition should be strictly governed by parity selection rules for the *gerade* and *ungerade* final core hole states, giving rise to purely odd and even angular momenta in the corresponding partial waves of

the $1\sigma_g$ and $1\sigma_u$ photolines. The well-known alternating intensities in rotational spectra resulting from the symmetry that must be imposed on the nuclear spin function to make the complete eigenfunction of the molecule either symmetric or antisymmetric could be viewed as the nuclear analogue of this electronic selectivity behaviour. Replacing one particle in such a system by a different one leads to a complete breakdown of the symmetry properties of the system.

The showcase example for such complete change in behaviour is the rotational structure in homonuclear diatomic molecules under isotope substitution mentioned above. Here the symmetry selection rules totally collapse and all forbidden or suppressed rotational transitions become equally allowed⁴. On the other hand, the electronic charge distribution in such a molecule is virtually unchanged by isotope substitution. Indeed, according to the Born–Oppenheimer approximation with its complete decoupling of nuclear and electronic motion, no change should occur in the electronic wavefunction of a hetero-isotopic homonuclear molecule. Known violations of the symmetry rules for the ground vibrational and electronic state of homonuclear diatomic molecules are miniscule⁵, and a symmetry breakdown has been observed for highly excited states only^{6–9}. Hence, any observable isotope effects on the electronic wavefunction for core electrons, the key element for chemical and structural analysis of matter¹⁰, might seem quite unexpected.

Here we show that inversion symmetry indeed causes non-local, coherent behaviour of the core electron photoemission from homonuclear diatomic molecules such as N₂ (ref. 11 and references therein). Our results show that this non-locality changes in a continuous way into partially localized behaviour, if inversion symmetry violations such as isotope substitution are induced.

The experiments were performed with vacuum ultraviolet synchrotron radiation from beamline BW3 of HASYLAB at DESY and beamline UE56/2-PGM1 and UE56/1-PGM at BESSY using a set of electron time-of-flight spectrometers in combination with an ion time-of-flight spectrometer with a position-sensitive anode (Fig. 1). This set-up makes it possible to determine all photoelectron and fragment ion momenta in coincidence, yielding, in the axial recoil approximation, the photoelectron angular distribution of fixed-in-space molecules^{12,13}. Because the N₂:N(1s)-doublet splitting of less than 100 meV (ref. 14) had to be resolved while data were acquired over several days, these measurements required extremely high-energy resolution of both the beamline (40 meV) and our set-up (60 meV) as well as a very high photon beam stability, particularly regarding the photon energy (10^{-5}). Unresolved spectra would provide the sum of

¹Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany. ²Max-Born-Institut für Nichtlineare Optik und Kurzzeitspektroskopie, 12489 Berlin, Germany.

³California Institute of Technology, Pasadena, California 91125, USA. [†]Present addresses: Medical Faculty, Physics Department, University of Rijeka, 51000 Rijeka, Croatia (S.C.); Steacie Institute for Molecular Sciences, National Research Council Canada, Ottawa, Ontario K1A 0R6, Canada (O.G.); Institut für Physikalische Chemie, Universität Würzburg, 97074 Würzburg, Germany (B.L.); Institute of Multidisciplinary Research for Advanced Materials, Tohoku University, Sendai 980-8577, Japan (T.L. & G.P.); Max-Planck-Institut für Physik komplexer Systeme, 01187 Dresden, Germany (B.Z.).

the *gerade* and *ungerade* photoemission channels^{15–19}, which displays no effect of nonlocality of the core electrons.

Figure 2 shows the resolved molecule-frame angular distributions of the $1\sigma_g$ and $1\sigma_u$ photoelectrons, together with the angular difference arising from the interfering molecular charge distributions. Less characteristic, but still symmetry-specific behaviour occurs also in the non-coincident laboratory-frame angular distribution characterized by the photoelectron angular distribution parameter β (ref. 20) shown in Fig. 3. The distinction is particularly pronounced in the region of the trapped *f*-wave resonance at 9 eV above the $N_2:N(1s)$ threshold because of the dominance of this *f*-partial continuum wave in only one (*gerade*) of the two photoelectron channels²¹. Note that the predicted crossover of the two *gerade* and *ungerade* angular distribution curves marks the onset of an oscillation driven by the spatial interference of the outgoing photoelectron.

We investigated the transition to the symmetry-broken system by comparing the naturally most abundant $^{14,14}N_2$ nitrogen molecule to two different isotopomers: singly substituted $^{14,15}N_2$ and doubly substituted $^{15,15}N_2$ (both 99% purity). Using the electron spectrometer in a non-coincident mode, that is, without detecting the corresponding fragment ions, we studied the effect of isotope substitution on the photoelectron spectrum. These effects are best illustrated in the ratio of the photoelectron spectra of normal and substituted nitrogen. Figure 4 shows the ratio between the $1s$ photoelectron spectra of normal $^{14,14}N_2$ and the isotopomer $^{14,15}N_2$ detected at the β -independent ‘magic angle’ θ_m (54.7° with respect to the electric vector of the ionizing radiation as shown in Fig. 3), where the measured photoelectron intensity is directly proportional to the partial cross-section²⁰, as well as at 0° where the β -dependence is largest (see Fig. 3). The experimental data (purple circles) are shown together with a model calculation of the vibrational effect due to the mass-dependence of the vibrational

constant (dashed black line in Fig. 4c and d), which causes nuclear-dynamics-dependent spectral changes.

We attribute the variation of the cross-sections and angular distributions beyond this behaviour to the breaking of inversion symmetry in the singly substituted species, which results in a partial localization of the core hole. Whereas the inversion symmetry of N_2 is preserved in the doubly substituted species $^{15,15}N_2$, the electronic wavefunction in $^{14,15}N_2$ is slightly modified owing to the broken symmetry of the singly substituted molecule, where the centre of symmetry r_{inv} of the electric charges midway between the two atoms no longer coincides with the centre of mass r_{cm} (ref. 8, 9). The wavefunctions in the molecule with broken inversion symmetry lose their character as parity eigenfunctions and can be described by linear combinations of the original *gerade* and *ungerade* wavefunctions. This mixing leads to greater similarity in the cross-sections and angular distributions, resulting in an isotope-induced effect in the range of a few per cent.

Two questions arise in this context: How may we understand the size of the observed effect, and why has it not been seen before in the photoelectron spectra of any homonuclear diatomic molecule? Both questions are closely related and require energy considerations related to the so-called diagonal and nondiagonal asymmetries in a bipolar system. (Here the terms ‘diagonal’ and ‘nondiagonal’ refer to contributions to the total hamiltonian of the system, which appear as diagonal and off-diagonal elements, respectively, when the origin of the coordinate system is equidistant from the two nuclei rather than at the centre of mass⁸.) In $^{14,15}N_2$, the centre of mass r_{cm} is shifted away from the inversion centre r_{inv} by 1.7% of the bond length, resulting in 3.5% asymmetric motion per nucleus due to the

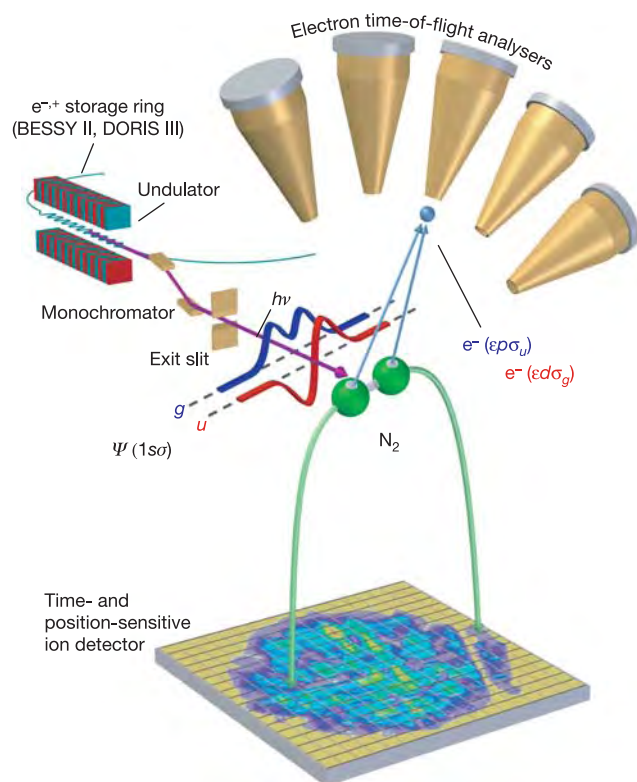


Figure 1 | Experimental set-up for a photoelectron-fragment ion coincidence experiment at a synchrotron radiation source. (For further details, see ref. 13.)

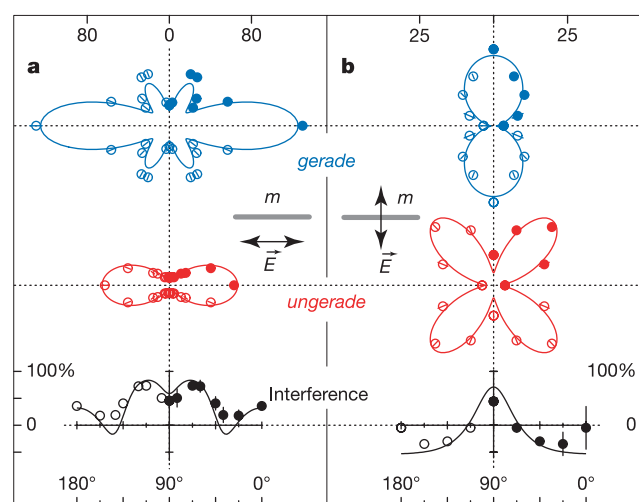


Figure 2 | Symmetry resolved photoelectron distributions in the molecule frame. Molecule frame photoelectron angular distributions (MPADs) for the *gerade* (upper panel) and *ungerade* (middle panel) $N(1s)$ core photoelectron emission of N_2 at a photon energy of $h\nu = 419$ eV measured in the plane perpendicular to the light propagation direction for molecules oriented parallel (a) and perpendicular (b) to the light polarization vector through selection by an ion momentum resolving imaging detector. The fractional interference angular pattern shown in the lower panel are the differences between the *gerade* and *ungerade* MPADs divided by their respective sums $(g - u)/(g + u)$. The open circles are the mirror images of the measured data points (full circles), which are obtained by a least-squares fit of the coincident spectra. The error bars reflect the statistical uncertainty (s.d.) of the fit. The solid lines are predictions for non-local, coherent electron emission calculated in the partially relaxed core Hartree-Fock (RCHF) approximation³¹ shown on a relative scale in arbitrary units marked at the upper margin, but unscaled with respect to each other. Note that the sensitivity of the measurements regarding the difference is considerably reduced at joint nodal points because of the low count rate in both transitions.

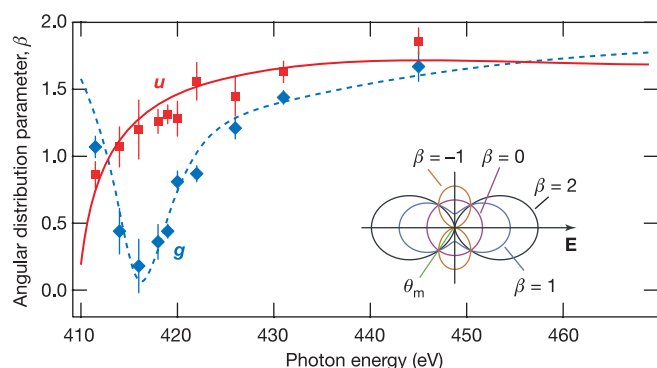


Figure 3 | Photoelectron angular distribution parameter β for $\text{N}_2:\text{N}(1s)$ electron emission in the photon-energy range 410–450 eV. The symbols represent the experimental data for the *gerade* (blue diamonds) and *ungerade* (red squares) state. The error bars reflect the statistical and calibrational error. The results are compared to calculations in the partially RCHF approximation³¹. The dashed line shows the results for the *gerade* state, whereas the solid line represents the calculation for the *ungerade* state. Our RCHF curves are generally in good agreement with the data obtained by the Kohn–Sham density functional theory³² and the random phase approximation³³ approach, in particular for the *ungerade* state (N. A. Cherepkov, personal communication).

relationship $r_{\text{cm}} = (m_{15} - m_{14}) / (m_{15} + m_{14}) \times r_{\text{inv}}$, where m_{14} and m_{15} are the masses of ^{14}N and ^{15}N , which determines the time during which the electronic wave packet experiences asymmetric motion of the nuclei with respect to the inversion centre. This is similar to the inverse effect of symmetry restoration by detuned excitation in resonant inelastic X-ray scattering²². In dissociating systems, a loss of symmetry has been observed for autoionization²³ and resonant Auger²⁴ lines owing to the Doppler shift during the emission process.

A system of identical non-overlapping (that is, strongly localized) particles gives rise to completely degenerate *gerade* and *ungerade* states. Their underlying symmetry character is inaccessible to

experimental exploration by photoelectron spectroscopy because the incoherent sums of *gerade* and *ungerade* states and of left- and right-hand states are, by definition, identical. A minimal delocalization resulting in a non-zero overlap between the two core orbitals is required to force a separation between the two symmetry-adapted states of the order of their natural lifetime width. Although this is just the case for $\text{N}_2:\text{N}(1s)$ photoemission, other core level photoelectron spectra of diatomic homonuclear molecules do not fulfil this requirement²⁵.

In contrast to core levels, the *gerade* and *ungerade* splitting for valence levels is very large (in the range of several electron volts) owing to the strong delocalization of most valence electrons. In fact, all ground states are of *gerade* symmetry—the corresponding *ungerade* state is unoccupied. One may view the core level splitting in N_2 as being caused by a core valence-coupling-induced tunnelling rate giving rise to a roughly 15% probability that the electrons from one site will be at the other atomic site²⁵. This tunnelling stabilizes the non-local, coherent character of the electronic state against asymmetric left/right, or, in our terminology, nondiagonal distortions such as a shift of the centre of mass away from the inversion centre, which, for small distortions, may be treated as a perturbation. With regard to localization, it is the ratio of the two associated energies that determines the size of isotope-induced effects in the photoelectron spectra of diatomic homonuclear molecules, analogous to the role of tunnelling versus correlation energy in a superfluid/Mott insulator transition²⁶.

Of all the energy effects caused by isotope substitution, we find that only one, the vibrational motion giving rise to an asymmetry energy of the order of several millielectronvolts, is of importance. All other effects, particularly hyperfine perturbations which cause *gerade*/*ungerade* symmetry breaking in highly excited states^{6,7} and isotope shifts inducing predissociation in isotopomers^{8,9}, are in the microelectronvolt range. The observed effect can therefore be viewed as the diatomic analogue of symmetry breaking by vibronic coupling in triatomic molecules²⁷ owing to the asymmetric part of the vibrational motion (in the range of 10 meV). Comparing this 10-meV fraction of the vibrational energy with the *gerade*/*ungerade* splitting of 100 meV yields a *gerade*/*ungerade* mixing coefficient

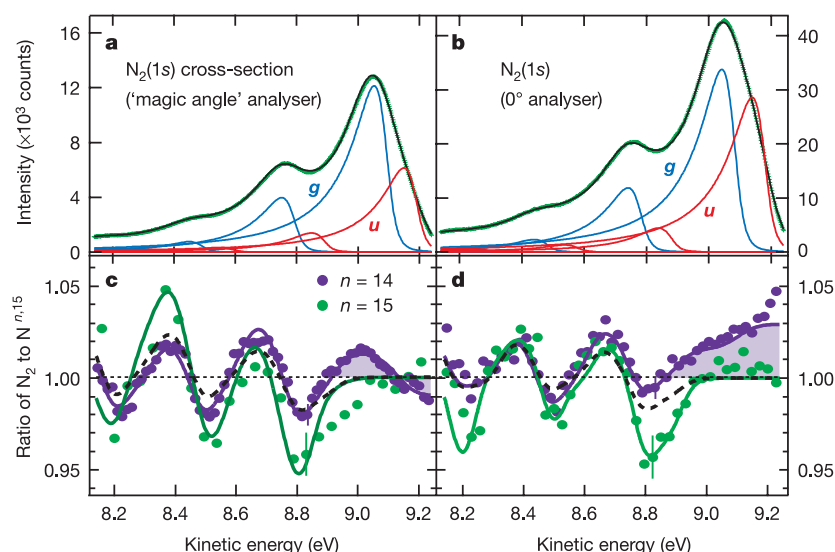


Figure 4 | Isotope effect on N_2 . **a, b**, High-resolution photoelectron spectrum of $^{14,14}\text{N}_2$ recorded at a photon energy of $h\nu = 419$ eV at the β -independent 'magic angle' θ_m (left) and at 0° with respect to the polarization vector of the ionizing radiation (right). The error bars indicate the typical statistical error. The measured spectrum (green line) in the range of the $1s$ photoline is shown together with the unconvoluted representation of a least-squares fit of its two symmetry-components *gerade* (blue) and

ungerade (red) and their respective vibrational progression up to the third vibrational level. **c, d**, Spectral ratio $^{14,14}\text{N}_2/^{14,15}\text{N}_2$ (purple) compared to $^{14,14}\text{N}_2/^{15,15}\text{N}_2$ (green) for the same angles as above. The solid lines are model calculations which include vibrational effects and, for $^{14,15}\text{N}_2$, the effect of symmetry breaking on the cross-sections. For $^{14,15}\text{N}_2$, the dashed line shows a model calculation of the vibrational effect only.

$[\Delta m/(m_{14} + m_{15})]/\Delta E_{gu}$ of 10%. To obtain the resulting relative change of the intensities in a first approximation, the square of the mixing coefficient must be multiplied by the normalized intensity difference of the *gerade* and *ungerade* channels $(I_g - I_u)/(I_g + I_u)$. At the photon energy considered here, this results in an estimated relative change of the cross-section of the order of 1%, which is consistent with the magnitude of the experimentally observed cross-section effect (Fig. 4c). The angle-dependent effect is enlarged owing to the role of the phase shifts between the photoelectron partial waves for all other emission directions besides the magic angle. It is worth mentioning that the size of the observed effect is still too small to be detected unambiguously in coincident measurements as shown in Fig. 2.

This analysis also explains why this effect has never been observed in valence photoionization, where the fractional size of the effect is more than two orders of magnitude smaller and where it must be measured on an absolute scale because there are no close-lying *gerade* and *ungerade* lines displaying effects in opposite directions. Such small absolute changes of less than 10^{-4} are inaccessible to photoelectron spectroscopy at present.

In summary, we have shown here that the inversion symmetry of a system indeed causes non-local, coherent behaviour of the otherwise localized core holes in homonuclear diatomic molecules such as N_2 . This non-locality of the electron emission and the remaining core hole is neither conserved nor completely destroyed by a distinct symmetry distortion such as isotope substitution, but instead changes in a continuous way into partially localized behaviour owing to the gradual breakdown of inversion symmetry, as reflected by the loss of interference and parity mixing of the outgoing photoelectron waves. This isotope effect on the electronic structure of a diatomic molecule, probed here by photoelectron spectroscopy, is the first experimentally observed effect of its kind, to our knowledge.

The continuous nature of this transition, of which we have seen just the onset, makes it possible to control the character of a quantum state from either local or non-local by applying distinct forces that either stabilize or destabilize the non-locality. This knowledge might be useful in other systems such as double quantum dots, which are envisaged as the future building blocks of quantum gates^{28,29}.

With the advent of free-electron lasers (FEL)³⁰, which will permit time-resolved pump-probe experiments in the vacuum ultraviolet region, new experiments will become feasible that can probe transitions between complete localization of the electrons on individual atomic sites and complete non-localization over identical sites in analogy to the studies of coherence reported in reference³. To this end, we envision an experiment where a nitrous oxide (N_2O) molecule is broken into an O and a N_2 fragment by an initial light pulse and the core photoionization of the N_2 fragment is then probed for various time delays as the oxygen moves further and further away. With increasing distance, the emission characteristics should change from the localized, incoherent case of N_2O —where the two nitrogen atoms are distinct because of the chemical shift induced by the oxygen—to the non-local, coherent case of N_2 .

METHODS

The model calculations shown in Fig. 4c and d are based on the results of a least-squares fit of the $^{14,15}N_2$ spectrum. The change of the vibrational constant is simulated by decreasing the energy spacing between the vibrational components by 5 meV, as predicted by a harmonic oscillator model. The symmetry-induced cross-section effect is modelled by further changing the intensity of the *gerade* and *ungerade* components. Changes to the Franck–Condon factors caused by the increased reduced mass of the substituted molecule are also included. However, they are found to be relatively small compared to the vibrational and cross-sectional effects. The increased mass leads to a decrease of the vibrational energy and therefore to pronounced oscillations in the intensity ratio, which coincide with the position of the vibrational progression. This vibrational effect should be more pronounced for the doubly substituted species $^{15,15}N_2$ owing to the larger

change of its vibrational energy resulting from its heavier mass. The model calculation reproduces those oscillations very well, but fails to explain the additional ‘wiggle’ (shaded area) at the high-energy end of the $N_2/^{14,15}N_2$ ratio, which is visible both in the ratio at the magic angle as well as at 0° with respect to the light polarization. This second effect, which does not appear in the $N_2/^{15,15}N_2$ ratio (shown in green), can only be explained by a change in both relative intensity as well as angular distribution of the *gerade* and *ungerade* components in $^{14,15}N_2$ compared to the two other isotopomers. The model calculation including these changes (solid lines) reproduces the experimental data even at the high-energy end of the $N_2:N(1s)$ photoline. It should be noted that the symmetry-induced effect is supposed to appear in all vibrational components; however, for the higher vibrational components below 8.9 eV, the symmetry effect tends to be masked by the vibrational effect. The pure symmetry effect is observed only in the lowest vibrational component above 9 eV.

Received 7 April; accepted 11 July 2005.

1. Lindner, F. et al. Attosecond double-slit experiment. *Phys. Rev. Lett.* **95**, 040401 (2005).
2. Briggs, J. S. & Walter, M. Oscillatory structure of molecular photoionisation cross-sections. *Phys. Essays* **13**, 297–302 (2000).
3. Hackermüller, L., Hornberger, K., Brezger, B., Zeilinger, A. & Arndt, M. Decoherence of matter waves by thermal emission of radiation. *Nature* **427**, 711–714 (2004).
4. Mulliken, R. S. Band spectra and atomic nuclei. *Trans. Faraday Soc.* **25**, 634–645 (1929).
5. Herzberg, G. *Molecular Spectra and Molecular Structure, Spectra of Diatomic Molecules* Vol. 1, 139 (Krieger, Malabar, Florida, 1989).
6. Pique, J. P., Hartmann, F., Bacis, R., Churassy, S. & Koffend, J. B. Hyperfine-induced *ungerade-gerade* symmetry breaking in a homonuclear diatomic molecule near a dissociation limit: $^{127}I_2$ at the $^2P_{3/2}-^2P_{3/2}$ limit. *Phys. Rev. Lett.* **52**, 267–270 (1984).
7. Critchley, A. D. J., Hughes, A. N. & McNab, I. R. Direct measurement of a pure rotation transition in H_2 . *Phys. Rev. Lett.* **86**, 1725–1728 (2001).
8. Cacciani, P. & Kokkoouline, V. Predissociation induced by *ungerade-gerade* symmetry breaking in $^6Li^7Li$ molecule. *Phys. Rev. Lett.* **84**, 5296–5299 (2000).
9. Bouloufa, N. et al. Predissociation induced by *ungerade-gerade* symmetry breaking in the $B^1\Pi_u$ state of the $^6Li^7Li$ molecule. *Phys. Rev. A* **63**, 042507 (2001).
10. Siegbahn, K., et al. *ESCA—Atomic, Molecular and Solid State Structure Studied by Means of Electron Spectroscopy* (Almqvist & Wiksells, Uppsala, 1967).
11. Broer, R. & Nieuwpoort, W. C. Hole localization and symmetry breaking. *J. Mol. Struct.* **458**, 19–25 (1999).
12. Heiser, F. et al. Demonstration of strong forward-backward asymmetry in the C1s photoelectron angular distribution from oriented CO molecules. *Phys. Rev. Lett.* **79**, 2435–2437 (1997).
13. Becker, U. Angle-resolved electron–electron and electron–ion coincidence spectroscopy: new tools for photoionization studies. *J. Electron. Spectrosc. Relat. Phenom.* **112**, 47–65 (2000).
14. Hergenbahn, U., Kugeler, O., Rüdell, A., Rennie, E. E. & Bradshaw, A. M. Symmetry-selective observation of the N 1s shape resonance in N_2 . *J. Phys. Chem. A* **105**, 5704–5708 (2001).
15. Shigemasa, E. et al. Angular distributions of $1s\sigma$ photoelectrons from fixed-in-space N_2 molecules. *Phys. Rev. Lett.* **74**, 359–362 (1995).
16. Pavlychev, A. A. et al. Dynamic properties of N and O $1s^{-1}\sigma_u^+$ shape resonances in N_2 and CO_2 molecules. *Phys. Rev. Lett.* **81**, 3623–3626 (1998).
17. Cherepkov, N. A. et al. Manifestation of many-electron correlations in photoionization of the K shell of N_2 . *Phys. Rev. Lett.* **84**, 250–253 (2000).
18. Jahnke, T. et al. Circular dichroism in K-shell ionization from fixed-in-space CO and N_2 molecules. *Phys. Rev. Lett.* **88**, 073002 (2002).
19. Weber, Th. et al. K-shell photoionization of CO and N_2 : is there a link between the photoelectron angular distribution and the molecular decay dynamics? *J. Phys. B* **34**, 3669–3678 (2001).
20. Yang, C. N. On the angular distribution in nuclear reactions and coincidence measurements. *Phys. Rev.* **74**, 764–772 (1948).
21. Dehmer, J. L. & Dill, D. Shape resonances in K-shell photoionization of diatomic molecules. *Phys. Rev. Lett.* **35**, 213–215 (1975).
22. Skytt, P. et al. Quenching of symmetry breaking in resonant inelastic X-ray scattering by detuned excitation. *Phys. Rev. Lett.* **77**, 5035–5038 (1996).
23. Golovin, A. V. et al. Observation of site-specific electron emission in the decay of superexcited O_2 . *Phys. Rev. Lett.* **79**, 4554–4557 (1997).
24. Björneholm, O. et al. Doppler splitting of in-flight auger decay of dissociating oxygen molecules: the localization of delocalized core holes. *Phys. Rev. Lett.* **84**, 2826–2829 (2000).
25. Kosugi, N. Spin-orbit and exchange interactions in molecular inner shell spectroscopy. *J. Electron. Spectrosc. Relat. Phenom.* **137–140**, 335–343 (2004).
26. Greiner, M., Mandel, O., Esslinger, T., Hänsch, T. W. & Bloch, I. Quantum

- phase transition from a superfluid to a Mott insulator in a gas of ultracold atoms. *Nature* **415**, 39–44 (2002).
27. Domcke, W. & Cederbaum, L. S. Vibronic coupling and symmetry breaking in core electron ionization. *Chem. Phys.* **25**, 189–196 (1977).
 28. Bayer, M. *et al.* Coupling and entangling of quantum states in quantum dot molecules. *Science* **291**, 451–453 (2001).
 29. Hayashi, T., Fujisawa, T., Cheong, H. D., Jeong, Y. H. & Hirayama, Y. Coherent manipulation of electronic states in a double quantum dot. *Phys. Rev. Lett.* **91**, 226804 (2003).
 30. Wabnitz, H. *et al.* Multiple ionization of atom clusters by intense soft X-rays from a free-electron laser. *Nature* **420**, 482–485 (2002).
 31. Zimmermann, B., Wang, K. & McKoy, V. Circular dichroism in K-shell ionization from fixed-in-space CO and N₂. *Phys. Rev. A* **67**, 042711 (2003).
 32. Stener, M., Fronzoni, G. & Decleva, P. Time dependent density functional study of the symmetry resolved N 1s photoionization in N₂. *Chem. Phys. Lett.* **351**, 469–474 (2002).
 33. Semenov, S. K. & Cherepkov, N. A. Generalization of atomic random-phase-approximation method for diatomic molecules. II. N₂ K-shell photoionization. *Phys. Rev. A* **66**, 022708 (2002).
- Acknowledgements** We thank J. Bozek, R. Díez Muiño, F. J. García de Abajo, C. S. Fadley and M. A. Van Hove for many discussions about core-hole delocalization and coherent photoelectron emission in homonuclear molecules. The assistance of R. Püttner and F. Gel'mukhanov in the interpretation of the effect of isotope substitution on the vibrational structure and the Franck-Condon factors is also acknowledged. The work was partly supported by the Bundesministerium für Bildung und Forschung (BMBF) and the Alexander von Humboldt Foundation (B.Z.).
- Author Information** Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to U.B. (becker_u@fhi-berlin.mpg.de).

LETTERS

A mesoporous germanium oxide with crystalline pore walls and its chiral derivative

Xiaodong Zou¹, Tony Conradsson¹, Miia Klingstedt¹, Mike S. Dadachov² & Michael O'Keeffe³

Microporous oxides are inorganic materials with wide applications in separations, ion exchange and catalysis^{1–3}. In such materials, an important determinant of pore size is the number of M (where M = Si, Ge and so on) atoms in the rings delineating the channels¹. The important faujasite structure exhibits 12-ring structures while those of zeolites^{4,5}, germanates^{6–8} and other⁸ materials can be much larger. Recent attention has focused on mesoporous materials with larger pores of nanometre scale^{9–11}; however, with the exception of an inorganic–organic hybrid¹², these have amorphous pore walls, limiting many applications. Chiral porous oxides are particularly desirable for enantioselective sorption and catalysis¹³. However, they are very rare in microporous^{14,15} and mesoporous¹⁶ materials. Here we describe a mesoporous germanium oxide, SU-M, with gyroidal channels separated by crystalline walls that lie about the G (gyroid) minimal surface as in the mesoporous MCM-48 (ref. 9). It has the largest primitive cell and lowest framework density of any inorganic material and channels that are defined by 30-rings. One of the two gyroidal channel systems of SU-M can be filled with additional oxide, resulting in a mesoporous crystal (SU-MB) with chiral channels.

We first describe the framework structure of SU-M, which was prepared by standard hydrothermal methods without using surfactants, but with an organic amine as the structure-directing agent, similar to conventional zeolite synthesis (see Methods). SU-M is cubic and has a unit cell of $a = 51.3 \text{ \AA}$. Similar to MCM-48, SU-M has symmetry $Ia\bar{3}d$ —the most complex cubic symmetry characterized by non-intersecting rotation axes, and glide rather than mirror planes—and structures with this symmetry are notoriously hard to illustrate¹⁷. The volume of the primitive cell is $67,640 \text{ \AA}^3$; searches of the Cambridge Crystallographic (<http://www.ccdc.cam.ac.uk>) and Inorganic Crystal (<http://icsdweb.fiz-karlsruhe.de>) Structure Databases found only one inorganic material, a molybdenum oxide cluster compound¹⁸, with a larger primitive cell.

SU-M is built from a unique $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ cluster (Fig. 1a) with O atoms singly coordinated to Ge corresponding to OH. The cluster consists of a central core of four octahedrally coordinated Ge atoms and six tetrahedrally coordinated Ge atoms. Each cluster is linked to five other clusters (Fig. 1b) via Ge–O–Ge bonds to form a three-dimensional framework with overall stoichiometry $\text{Ge}_{10}\text{O}_{20.5}(\text{OH})_3$. There are 96 $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ clusters per unit cell that build crystalline walls about the G minimal surface, and that correspond to the amorphous walls in MCM-48⁹. A (111) slab of the structure shown in Fig. 1c demonstrates a complex system of linked $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ clusters forming big cavities ($>20 \text{ \AA}$, see Fig. 1c). The big cavities are at positions 16b (with coordinates $1/8, 1/8, 1/8$ and their symmetry equivalents). Each cavity is connected to three other cavities through windows of 30 $\text{GeO}_4/\text{GeO}_6$ polyhedra (30-rings) (Fig. 1d) to form giant gyroidal channels. SU-M contains two such channels of

opposite chirality (see Supplementary Video 1); the largest opening between the two channels is a 12-ring formed by six $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ clusters that are located around the positions 16a (the origin of the unit cell and its symmetry equivalents) (Fig. 1c).

The gyroidal channels can be described as three-coordinated nets with vertices at the centres of the big cavities and edges connecting the nearest cavities of the channels, as shown in Fig. 2a. Each channel forms such a three-coordinated net, well known as the net of the Si

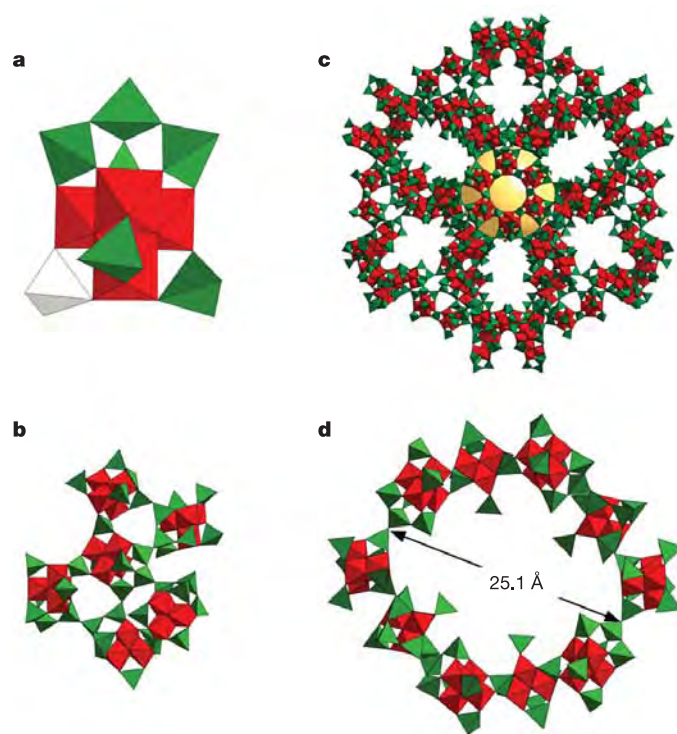


Figure 1 | Linkage of $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ clusters in SU-M. **a**, The $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ cluster built from six GeO_4 tetrahedra (green) and four GeO_6 octahedra (red). The white tetrahedron belongs to an adjacent cluster. **b**, A $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ cluster as in **a** linked to five neighbouring clusters. **c**, A 30-Å-wide slab with a big cavity at the centre. The yellow ball represents an oblate spheroid at the centre of the cavity that does not touch the centre of any framework atom. It has an equatorial diameter of $26.2 \text{ \AA}$ and a polar diameter of $18.6 \text{ \AA}$ and a volume equal to that of a sphere of diameter $23.4 \text{ \AA}$. **d**, A 30-ring window formed by ten $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ clusters. The free diameter of the 30-ring is $10.0 \times 22.4 \text{ \AA}$, assuming the van der Waals diameter of oxygen $2.7 \text{ \AA}$. The big cavity at the centre in **c** is connected to three other big cavities (upper-left, upper-right and below) through the 30-ring windows.

¹Structural Chemistry, Stockholm University, SE-106 91 Stockholm, Sweden. ²Corpuscular Inc., 52 Papania Drive, Mahopac, New York 10541, USA. ³Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona 85287, USA.

atoms in the SrSi_2 structure, with symbol **srs** (for a database of nets, see ref. 19). The repeat unit ('tile') of the **srs** net²⁰ is composed of three 10-rings (the red unit in Fig. 2b). This net is one of the five regular three-periodic nets²⁰; it is the only chiral one (with space group $I4_132$). The **srs** net has the property that it can intergrow with its enantiomorph in such a way that all the 10-rings of one net are catenated with the 10-rings of the other. The combined structure has symmetry $Ia\bar{3}d$ (Fig. 2a). Now imagine the nets uniformly inflated as suggested in Fig. 2c, until they meet at a surface. This continuous periodic surface of negative curvature is known as the G (gyroid) minimal surface²¹ and is the underlying structure of the walls of mesoporous materials SU-M and MCM-48 (ref. 9).

The crystalline wall of SU-M can be described as a 5-coordinated net with vertices at the centres of the $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ clusters. The net is a two-dimensional (2D) tiling of the G surface, with vertex symbol $3^2.4.3.6$, forming an infinite polyhedron²², as shown in Fig. 2e. This net is known as **fcz**¹⁹ and can be described also in terms of a 3D tiling of space²³ by tiles (Fig. 2f): a 'small' tile with face symbol²⁰ $[3^8.4^2.10^2]$ (Fig. 3a) and a 'big' tile with face symbol $[6^2.10^3]$ (Fig. 3b). The big tiles (16 per unit cell), with symmetry $32 (D_3)$ are centred at the vertices of the two interpenetrating **srs** nets, and correspond to

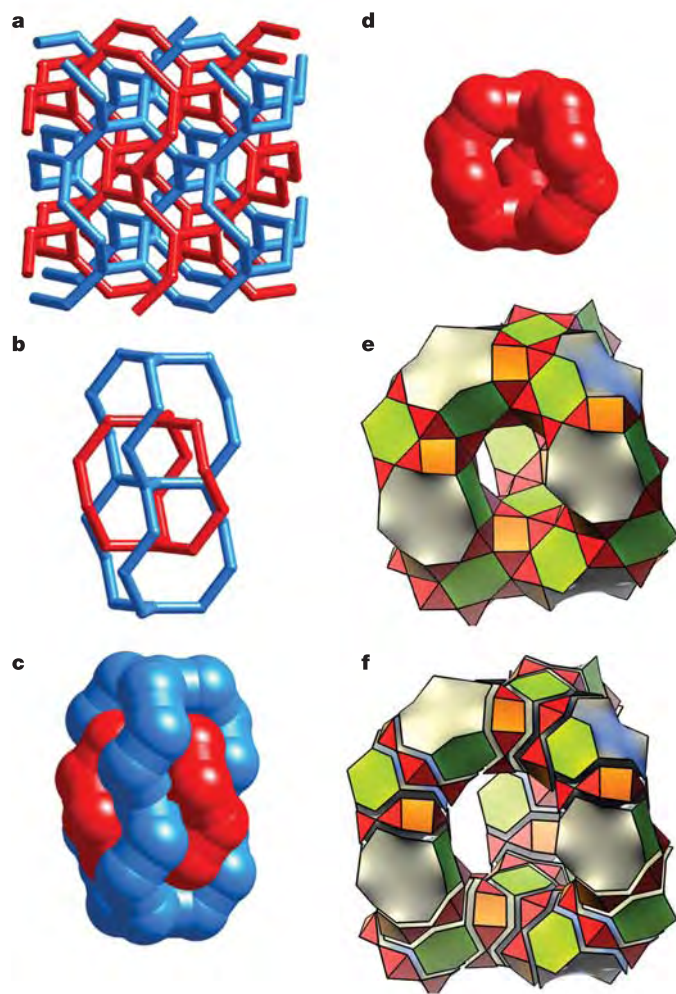


Figure 2 | Hierarchical description of the fcz net¹⁹. **a**, Two interpenetrating **srs**²⁰ nets—positions $16b$ of $Ia\bar{3}d$. **b**, A fragment of **a**. The red unit outlines a tile²⁰ for one of the nets. **c**, The same as **b** but inflated so the two parts meet at a common surface, the G surface. **d**, The inflated red unit alone. **e**, The net **fcz** as an infinite polyhedron forming a 2D tiling $3^2.4.3.6$ of the G surface. **f**, The same fragment of the structure shown as an exploded 3D tiling of space by big and small tiles. Notice that not all the tiles are shown—the full set of tiles completely fills space.

Table 1 | Comparison of the faujasite, SU-12 and SU-M structures

Structure	Net ²⁰	Cluster	Framework density (atoms nm ⁻³)	Largest ring	Available V	Occupiable V
Faujasite ²⁵	fau	Si	13.45	12	0.544	0.295
SU-12 ⁸	fee	Ge ₇	8.58	24	0.701	0.411
SU-M	fcz	Ge ₁₀	7.10	30	0.746	0.508

'Cluster' is the number of metal atoms per vertex of the underlying net. Framework density and volumes (V) are reported as fractions of the total volume. 'Available V' is the volume not occupied by spheres with van der Waals radius centred on the framework atoms.

'Occupiable V' was calculated for a probe sphere of radius 1.5 Å exploring the framework as described by Connelly²⁹ and implemented in the Cerius-2 computer program.

the big cavities in SU-M (Fig. 1c and 3d). The small tiles are at the centres of the links and correspond to the 30-ring windows of SU-M (Fig. 1d and 3c). The structure of SU-M can be built by replacing the vertices of the **fcz** net with $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ clusters. This last step is an example of 'scale chemistry'²⁴, as illustrated in Fig. 4.

SU-MB is a chiral derivative of SU-M, prepared in the presence of hydrofluoric acid (see Methods below). In the structure of SU-MB, one half of the big tiles (big cavities) are filled with additional (Ge, O, F) clusters. These clusters, formulated as $\text{Ge}_7\text{O}_{16}\text{F}_3$, are familiar from other germanium oxide frameworks²⁵ including ASU-16 (ref. 6) and SU-12 (ref. 8). Six of these clusters (144 additional atoms, Fig. 5b) fit inside one big tile (big cavity), with three of them connected to each 12-ring window of the main framework through the terminal atoms of the $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ clusters (Fig. 5a). The occupied big tile and its contents (that is the unit shown in Fig. 5a) now has composition $\text{Ge}_{222}\text{X}_{516}$ (where X = O, OH or F). The most remarkable aspect of SU-MB is, however, the fact that only half of the cavities are filled (see Supplementary Video 1), specifically all those of one hand, and the symmetry is reduced to $I4_132$. The system of empty pores and channels is accordingly chiral and corresponds to one of the red or blue nets of Fig. 2a: that is, the topology of the chiral net **srs**²⁰. We note that a chiral zeolite structure, UCSB-7, with a similar pore system has been reported¹⁴; however, in UCSB-7 the chirality is induced by ordering of the framework atoms, rather than by blocking

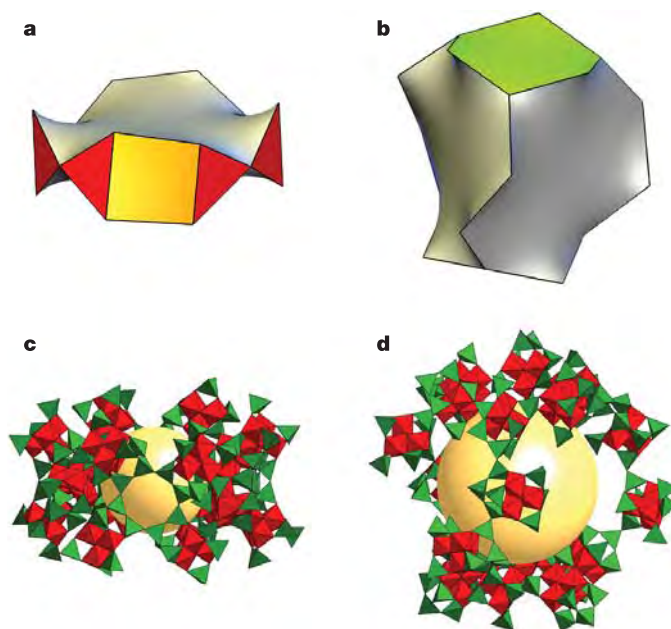


Figure 3 | The tiles of SU-M. **a**, A small tile of Fig. 2 and **c** the same tile in SU-M with each original vertex decorated with $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ clusters. The shortest ring around the perimeter in **c** involves 30 Ge atoms (see Fig. 1). **b**, A big tile of Fig. 2; **d**, the same tile in SU-M with the vertices similarly decorated. The yellow balls correspond to the largest sphere that fits inside each tile. The radius of the ball is 13.1 Å in **c** and 18.6 Å in **d**.

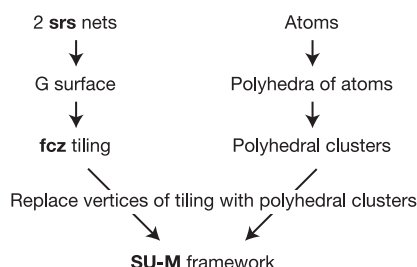


Figure 4 | The hierarchical nature of the structure of the SU-M framework.

one set of channels. We note also that the unit-cell volume of SU-MB ($a = 50.8 \text{ \AA}$) is over twenty times that of UCSB-7 ($a = 18.6 \text{ \AA}$), and the pores are much larger. As many as 336 (Ge, O, F) polyhedra can be incorporated into one of the gyroidal channels per unit cell. The chiral mesoporous silica¹⁶ is also very different as it consists of rifled parallel channels arranged as in MCM-41. At this point we are unable to account for the formation of chiral material in SU-MB and note that it will be a significant challenge to produce enantiopure material, presumably using chiral templates in the synthesis.

The germanium oxide framework is charged, with a formal charge of -4 per $\text{Ge}_{10}\text{O}_{24}(\text{OH})_3$ cluster in SU-M and an additional charge of -3 per $\text{Ge}_7\text{O}_{16}\text{F}_3$ cluster in SU-MB. The counter charge is provided by a protonated aliphatic diamine (2-methylpentamethylenediamine, MPMD), although in the crystal structures these amines and included water are incompletely resolved because of disorder. The chemical composition of SU-M is accordingly written $[(\text{H}_2\text{MPMD})_2(\text{H}_2\text{O})_x][\text{Ge}_{10}\text{O}_{20.5}(\text{OH})_3]$. For SU-MB the corresponding formulation is $[(\text{H}_2\text{MPMD})_{5.5}(\text{H}_2\text{O})_x][\{\text{Ge}_{10}\text{O}_{21}(\text{OH})_2\}_2 \cdot [\text{Ge}_7\text{O}_{14}\text{F}_3]]$.

Given the large scale of SU-M, it is interesting to compare it with other low-density materials. The faujasite framework²⁶ (Fig. 5) is often cited as the paradigm of a low-density zeolite framework. ASU-16 (ref. 6), specifically in its lower density conformation SU-12 (ref. 8), had the previous record for low framework density for an oxide material. Some properties related to density and porosity are listed for these three materials in Table 1. We note that in SU-M over 50% of the total volume is accessible to a probe sphere of radius $1.5 \text{ \AA}$ (appropriate for a molecule like water), in contrast to faujasite, in which less than 30% of the total volume is accessible.

Despite the presence of counterions and solvent, we have shown in preliminary experiments on as-synthesized SU-M that the material has permanent porosity, as shown by the observation of reversible type-I nitrogen adsorption isotherm, with a pore diameter of $12 \text{ \AA}$ deduced by non-local density functional theory²⁷ (see Supplementary Fig. 1). The BET (Brunauer–Emmett–Teller) surface area was $214 \text{ m}^2 \text{ g}^{-1}$ (note that because of the larger atomic weight of Ge compared to Si this would correspond to $368 \text{ m}^2 \text{ g}^{-1}$ for a silicate). One might also expect significant increase when post-synthetic treatment of the material is optimized. Energy dispersive spectroscopy showed that the MPMD cations in SU-M could be completely exchanged by Cs^+ , K^+ and Na^+ . The structure of the ion-exchanged SU-M was maintained, but with a slightly smaller unit cell, as shown by both single-crystal X-ray diffraction and X-ray powder diffraction. The crystallinity of SU-M was maintained when heated up to 320°C in air, indicated by the *in situ* X-ray powder diffraction (see Supplementary Fig. 2). Thermogravimetric analysis of SU-M in N_2 showed three steps of weight loss (see Supplementary Fig. 3), corresponding to surface water (10%, $20\text{--}100^\circ\text{C}$), crystal water (8%, $150\text{--}250^\circ\text{C}$) and partial decomposition of the protonated MPMD (6%, $320\text{--}450^\circ\text{C}$).

A mesoporous GeO_2 has recently been made by a surfactant templating method²⁸; it has a thermal stability very similar to that of SU-M, but less than that of silica-based materials. However, we

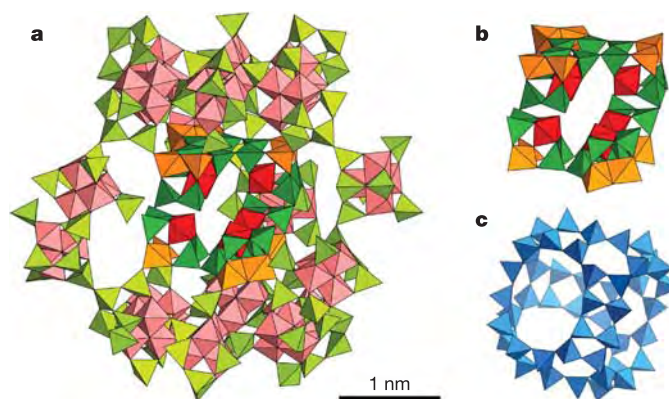


Figure 5 | Cavities in SU-MB and faujasite. **a**, The filled big tile in SU-MB. **b**, The six $\text{Ge}_7\text{O}_{16}\text{F}_3$ clusters in the interior of **a**; orange polyhedra are trigonal bipyramids. **c**, A faujasite supercage on the same scale. Note the 1-nm scale marker.

have previously shown⁸ that replacement of tetrahedral Ge in Ge cluster oxides by Si can significantly enhance stability, and this is a strategy worth pursuing for SU-M as well. This mesoporous material has a hexagonal array of parallel channels of the MCM-41 type but is otherwise amorphous. In contrast, the preparation of SU-M (see Methods) involves no surfactant. We believe it is significant that we have also found (work to be published elsewhere) that an ordered crystalline germanium oxide of the MCM-41 type constructed from clusters very similar to those in SU-M can also be made by a surfactant-free process. These observations open up the possibility that there may be more general routes to ordered mesoporous materials, including ones with chiral channel systems, of which the ones reported here are just the first.

METHODS

Synthesis. Both SU-M and SU-MB were synthesized under hydrothermal conditions from a homogenous solution of germanium dioxide, MPMD and water with the molar ratios of 1:8–10:38–40. In addition, hydrofluoric acid with a molar ratio of $\text{GeO}_2\text{:HF} = 1\text{:}1.5$ was added for the synthesis of SU-MB. The solutions were heated at 165°C in Teflon-lined Parr autoclaves under autogenous pressure and the synthesis time was seven days for SU-M and 11 days for SU-MB. Octahedral crystals, with sizes of $160 \times 160 \times 160 \mu\text{m}^3$ for SU-M and $80 \times 80 \times 80 \mu\text{m}^3$ for SU-MB, were obtained (see Supplementary Fig. 4).

Crystallographic studies. X-ray diffraction data were collected at 170 K on a STOE IPDS diffractometer equipped with an image plate and graphite-monochromatized $\text{MoK}\alpha$ radiation ($\lambda = 0.71073 \text{ \AA}$) from a rotating anode. The structures were solved by direct methods and refined by full-matrix least-squares techniques against F^2 .

For SU-M, 134,976 reflections, of which 8,922 were unique, were collected in the region $4.34^\circ < 2\theta < 48.18^\circ$. The space group is $Ia\bar{3}d$ and the unit-cell dimensions are $a = 51.335(3) \text{ \AA}$, $Z = 96$, $V = 135,282(14) \text{ \AA}^3$. All framework atoms, all nitrogen and more than half of the carbon from MPMD cations were located. Several water oxygen atoms were also located. All non-hydrogen framework atoms were refined anisotropically. $R_1 = 0.0744$ for 6,519 reflections with $I > 2\sigma(I)$ and 0.1057 for all 8,922 reflections; $wR_2 = 0.2362$ and S (the goodness of fit on F^2) = 1.013. Crystal data and details of structure determination are given in Supplementary Table 1. Atomic coordinates and equivalent isotropic displacement parameters are given in Supplementary Table 2.

For SU-MB, 34,928 reflections, of which 11,888 were unique, were collected in the region $7.94^\circ < 2\theta < 42.98^\circ$. The space group is $I4_132$ and the unit-cell dimensions are $a = 50.873(3) \text{ \AA}$, $Z = 48$, $V = 131,662(13) \text{ \AA}^3$. Owing to the relatively small crystal size, more than half of the reflections have intensities less than $2\sigma(I)$. The structure was solved by direct methods. All framework atoms and some of the nitrogen and carbon from MPMD cations were located. Several water oxygen atoms were also located. Only the germanium atoms were refined anisotropically. $R_1 = 0.0858$ for 6,403 reflections with $I > 2\sigma(I)$ and 0.1678 for all 11,888 reflections; $wR_2 = 0.2309$ and $S = 0.966$. Crystal data and details of

structure determination are given in Supplementary Table 3. Atomic coordinates and equivalent isotropic displacement parameters are given in Supplementary Table 4.

In situ X-ray powder diffraction. *In situ* X-ray powder diffraction was performed on a Huber Guinier camera 670 equipped with an imaging plate, using synchrotron radiation at Beamline I711, Max-lab, Lund University, Sweden. The as-synthesized SU-M was heated in air from 20 to 340 °C with an average heating rate of 7 °C min⁻¹. X-ray powder diffraction data were collected at every 50 °C up to 200 °C and every 20 °C from 200 to 340 °C.

Ion exchange. The ion exchange was performed in solutions of CsCl (1 M), KCl (1 M) and NaCl (1 M) at 20 °C for 20 h. The ion-exchanged crystals were first washed with water, then ethanol, and finally dried at 80 °C for 2 h.

Adsorption study. Nitrogen adsorption and desorption isotherm of as synthesized SU-M was measured at 77 K on a Micromeritics ASAP 2020 system. The sample was degassed first at 297 K for 5 h and then at 357 K for 5 h.

Received 26 April; accepted 28 July 2005.

- Davis, M. E. Ordered porous materials for emerging applications. *Nature* **417**, 813–821 (2002).
- Cheetham, A. K., Férey, G. & Loiseau, T. Open-framework inorganic materials. *Angew. Chem. Int. Edn Engl.* **38**, 3269–3292 (1999).
- van Bekkum, H., Jacobs, P. A., Flanigen, E. M. & Jansen, J. C. (eds) *Introduction to Zeolite Science and Practice* 2nd edn (Elsevier, New York, 2001).
- Davis, M. E., Saldarriaga, C., Montes, C., Garces, C. & Crowder, C. A molecular sieve with 18-membered rings. *Nature* **331**, 698–699 (1988).
- Estermann, M., McCusker, L. B., Baerlocher, Ch., Merrouche, A. & Kessler, H. A synthetic gallophosphate molecular-sieve with a 20-tetrahedral-atom pore opening. *Nature* **352**, 320–323 (1991).
- Plévert, J. *et al.* A flexible germanate structure containing 24-ring channels and with a very low framework density. *J. Am. Chem. Soc.* **123**, 12706–12707 (2001).
- Zhou, Y. *et al.* A large 24-membered-ring germanate zeolite-type open-framework structure with three-dimensional intersecting channels. *Angew. Chem. Int. Edn Engl.* **40**, 2166–2168 (2001).
- Tang, L., Dadachov, M. S. & Zou, X. D. SU-12: a silicon-substituted ASU-16 with circular 24-rings and templated by a monamine. *Chem. Mater.* **17**, 2530–2536 (2005).
- Kresge, C. T., Leonowicz, M. E., Roth, W. J., Vartuli, J. C. & Beck, J. S. Ordered mesoporous molecular sieves synthesized by a liquid crystal template mechanism. *Nature* **359**, 710–712 (1992).
- Zhao, D. *et al.* Triblock copolymer synthesis of mesoporous silica with periodic 50 to 300 Å pores. *Science* **279**, 548–552 (1998).
- Terasaki, O. (ed.) *Mesoporous and Related Nano-structured Materials* (Elsevier, New York, 2004).
- Inagaki, S., Guan, S., Ohsuna, T. & Terasaki, O. An ordered mesoporous organosilica hybrid material with crystal-like wall structure. *Nature* **416**, 304–307 (2002).
- Baiker, A. Chiral catalysis on solids. *Curr. Opin. Solid State Mater. Sci.* **3**, 86–93 (1998).
- Gier, T. E., Bu, X., Feng, P. & Stucky, G. D. Synthesis and organization of zeolite-like materials with three-dimensional helical pores. *Nature* **395**, 154–157 (1998).
- Wang, Y., Yu, J., Guo, M. & Xu, R. [$\{Zn_2(HPO_4)_4\}\{Co(dien)_2\}H_3O$]: A zinc phosphate with multidirectional intersecting helical channels. *Angew. Chem. Int. Edn Engl.* **42**, 4089–4092 (2003).
- Che, S. *et al.* Synthesis and characterization of chiral mesoporous silica. *Nature* **429**, 281–284 (2004).
- Andersson, S. & O'Keeffe, M. Body-centered cubic cylinder packing and the garnet structure. *Nature* **267**, 605–606 (1976).
- Müller, A., Koop, M., Bögge, H., Schmidtman, M. & Beugholt, C. Exchanged ligands on the surface of a giant cluster: $[(MoO_3)_{176}(H_2O)_{63}(CH_3OH)_{17}H_n]^{(32-n)-}$. *Chem. Commun.* 1501–1502 (1998).
- O'Keeffe, M., Yaghi, O. M., Moler, D., Joshi, G., Ockwig, N. & Delgado-Friedrichs, O. *Reticular Chemistry Structure Resource* <http://okeeffe-ws1.la.asu.edu/RCSR/home.htm> (Arizona State Univ., Arizona, 2004).
- Delgado-Friedrichs, O., O'Keeffe, M. & Yaghi, O. M. Three-periodic nets and tilings: regular and quasiregular nets. *Acta Crystallogr. A* **59**, 22–27 (2003).
- Andersson, S., Hyde, S. T., Larsson, K. & Lidin, S. Minimal surfaces and structures: from inorganic and metal crystals to cell membranes and biopolymers. *Chem. Rev.* **88**, 221–242 (1988).
- O'Keeffe, M. & Hyde, B. G. *Crystal Structures I: Patterns and Symmetry* 289–380 (Mineralogical Society of America, Washington DC, 1996).
- O'Keeffe, M. Tiling by numbers. *Nature* **400**, 617–618 (1999).
- Férey, G. Building units design and scale chemistry. *J. Solid State Chem.* **152**, 37–48 (2000).
- Plévert, J., Gentz, T. M., Groy, T. L., O'Keeffe, M. & Yaghi, O. M. Layered structures constructed from new linkages of $Ge_7(O,OH,F)_{19}$ clusters. *Chem. Mater.* **15**, 714–718 (2003).
- Hriljac, J. A., Eddy, M. M., Cheetham, A. K., Donohue, J. A. & Ray, G. J. Powder neutron diffraction and ^{29}Si MAS NMR studies of siliceous zeolite-Y. *J. Solid State Chem.* **106**, 66–72 (1993).
- Ravikovitch, P. I., Haller, G. L. & Neimark, A. V. Density functional theory model for calculating pore size distributions: pore structure of nanoporous catalysts. *Adv. Colloid Interf. Sci.* **76–77**, 203–226 (1998).
- Lu, Q., Gao, F., Li, Y., Zhou, Y. & Zhao, D. Synthesis of germanium oxide mesostructures with a new intermediate state. *Micropor. Mesopor. Mater.* **56**, 219–225 (2002).
- Connolly, M. L. Solvent accessible surfaces of proteins and nucleic acids. *Science* **221**, 709–713 (1983).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Lidin for help and discussions, A. Chizmeshya for the volume data reported in Table 1, K. E. Christensen for checking ion-exchanged samples and A. Garcia-Bennett for the N₂ isotherm. L. Q. Tang and E. Karlsson participated in the synthesis. The project is supported by the Swedish Science Research Council. M.O'K. acknowledges support from the US National Science Foundation. X.D.Z. is a Research Fellow of the Royal Swedish Academy of Sciences, supported by a grant from the Alice and Knut Wallenberg Foundation.

Author Information The additional crystallographic data for SU-M (CCDC-278829) and SU-MB (CCDC-278830) can be obtained free of charge from The Cambridge Crystallographic Data Centre via http://www.ccdc.cam.ac.uk/data_request/cif. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to X.D.Z. (zou@struc.su.se).

LETTERS

Field evidence for surface-wave-induced instability of sand dunes

Hicham Elbelrhiti^{1,2}, Philippe Claudin¹ & Bruno Andreotti¹

Field studies of barchans—crescent-shaped dunes that propagate over solid ground under conditions of unidirectional wind¹—have long focused on the investigation of an equilibrium between sand transport by wind and the control of air flow by dune topography^{2–4}, which are thought to control dune morphology and kinematics^{5–7}. Because of the long timescale involved, however, the underlying dynamic processes responsible for the evolution of dune fields remain poorly understood⁸. Here we combine data from a three-year field study in the Moroccan Sahara with a model study to show that barchans are fundamentally unstable and do not necessarily behave like stable solitary waves, as suggested previously^{9–12}. We find that dune collisions and changes in wind direction destabilize the dunes and generate surface waves on the barchans. Because the resulting surface waves propagate at a higher speed than the dunes themselves, they can produce a series of new barchans of elementary size by breaking the horns of large dunes. The creation of these new dunes provides a mechanism for sand loss that prevents dune fields from merging into a single giant dune and therefore plays a fundamental role in the control of size selection and the development of dune patterns.

Very few barchans in a dune field exhibit the smooth crescent shapes that are simulated in models (Fig. 1e); instead they display more complex substructures. The windward slope and flanks of barchans generally present superimposed bed-forms which can become high enough to induce air-flow separation and thus secondary avalanche slip faces (Fig. 1b–d). On the basis of the numerical findings that two colliding barchans can cross through one another while still preserving their shape, these phenomena have been interpreted as small dunes climbing onto large ones^{9–11}. This is contradicted by our direct field investigation, during which we have followed the birth, growth, propagation and further evolution of these structures. We studied more than a hundred barchans in the region between Tarfaya (27° 56' N, –12° 56' W), Sidi Aghfinir (28° 06' N, –12° 03' W) and Lâayoune (27° 10' N, –13° 14' W), where the wind regime (wind rose on Fig. 1) is one of the most unimodal¹³. Two situations under which the dune surface becomes destabilized are identified: changes of wind direction (Fig. 1a–d) and collisions (Fig. 1f–i). We have precisely investigated the detailed nature of the unstable modes on five dunes displaying well-defined patterns. With the help of fixed markers, we observed that the undulations propagate downwind on the stoss (windward) slope and the flanks of the dune at a velocity $c \approx 2 \text{ m day}^{-1}$, which is typically ten times larger than that of the dune itself. Their wavelength and amplitude do not vary much in the course of their motion. We have measured the variations of height δh and sand flux δq on a cut line along a barchan horn (Fig. 2b). These quantities are proportional, which demonstrates that these undulations behave as plane propagating waves.

The nucleation and propagation of such waves on a sand bed is governed by the interaction between the bed profile, which modifies the air flow, and the sand transport, which controls the erosion and deposition processes. Along the upwind side of a hump, the streamlines converge, yielding an increasing wind and thus an increasing flux, so that erosion takes place^{2–4}. Conversely, the flux decreases on the downwind side causing accretion, which in total means that the bump translates downwind. However, the accretion does not start precisely at the crest but is shifted upwind, so that the bump is amplified. This instability mechanism is directly related to the asymmetry of the wind flow, which originates in the nonlinear inertial term of the Navier–Stokes equations. So far in this description, no length scale is involved, because the atmospheric boundary layer is fully turbulent, and the mechanism therefore predicts an unconditional instability at all wavelengths. There is however a small-scale cut-off for the instability related to the transient approach to saturation of the sand flux. As exemplified in Fig. 3b, the flux reaches its equilibrium value, determined by the wind strength, over a characteristic distance L called the saturation length. This effect shifts downwind the position at which the flux q is maximum, and thus stabilizes small bumps.

Previously⁸, we have derived a simple model (called⁸ C_C^c) accounting for all the above mechanisms: mass conservation, shape-flux coupling and flux saturation. Here we present field measurements of the quantities characterizing the unstable waves—wavelength, propagation speed and amplitude—and quantitatively compare them to the predictions of the model. Three dimensionless parameters A , B and D enter the model, which respectively govern the wind acceleration on a bump, the displacement upwind of the maximum velocity and the lateral coupling. Their values are tuned to reproduce the relations between the morphological parameters of barchans: height, width and length (Fig. 1e). The only timescale of the problem is related to the saturated flux Q on a flat bed and encodes the wind strength (wind roses on Fig. 1). The only characteristic length scale¹⁴ is the saturation length L introduced above, and we designed a specific experiment to measure it on a 20-m-long and 3-m-wide flat sand sheet prepared with a bulldozer (Fig. 3b). L is directly related to the length needed for a grain to reach the wind velocity, which scales¹⁵ with the grain density to fluid density ratio times the grain diameter d . Using the measured values $L = 1.7 \text{ m}$ and $d = 180 \mu\text{m}$, we obtain:

$$L \approx 4 \frac{\rho_{\text{sand}}}{\rho_{\text{air}}} d \quad (1)$$

in close agreement with our re-analysis of Bagnold's experiment¹ ($L = 2.3 \text{ m}$ for $d = 250 \mu\text{m}$).

To derive the dispersion relation of surface waves, we have examined large dunes for which the scale separation is sufficient for us to consider that they are flat at the scale of the waves. In the

¹Laboratoire de Physique et Mécanique des Milieux Hétérogènes (LPMHM), UMR CNRS 7636, 10 rue Vauquelin 75005 Paris, France. ²Geology Department, Université Ibn Zohr, BP 28/5, Cité Dakhla, 80000 Agadir, Morocco.

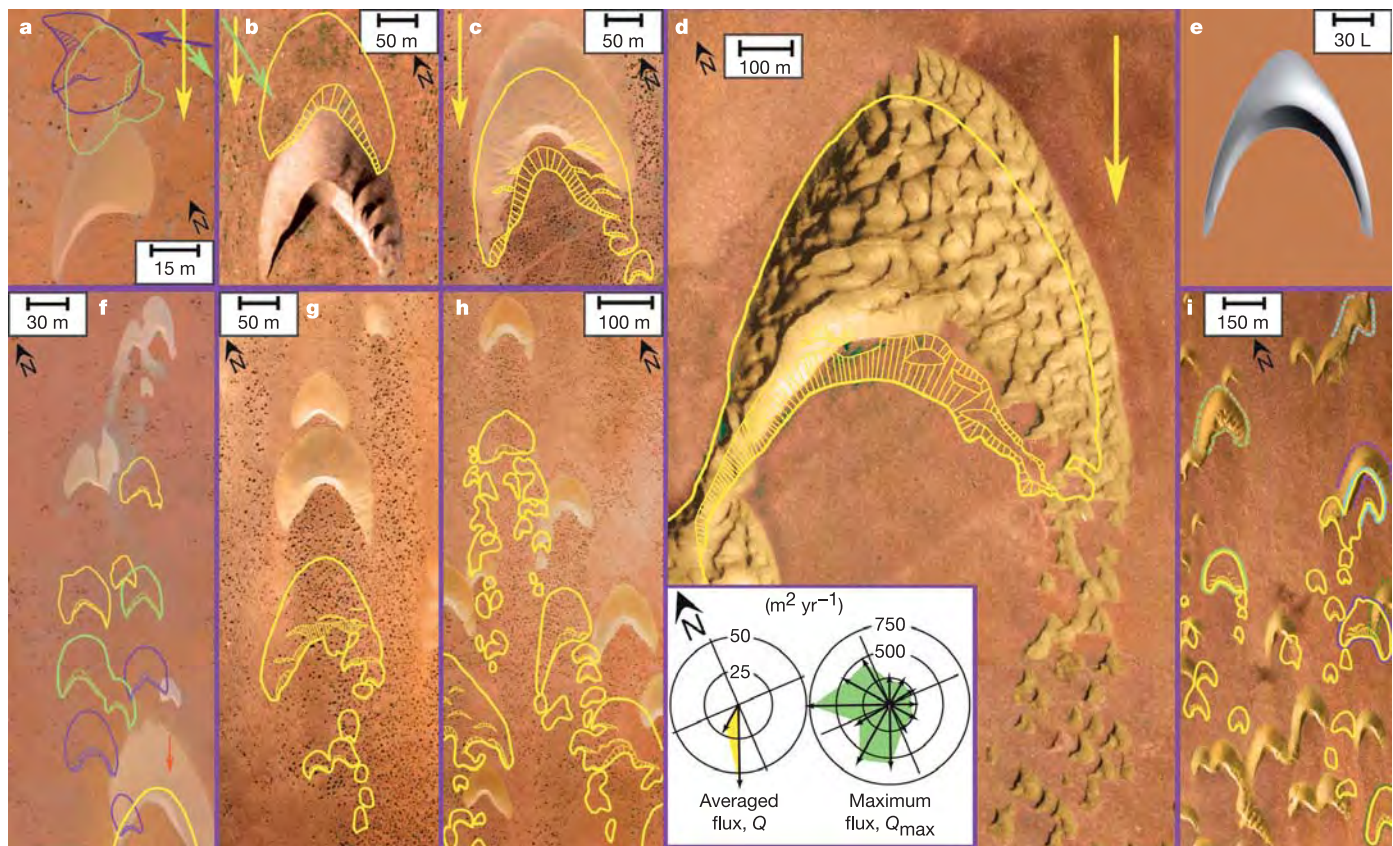


Figure 1 | Response of barchan dunes to changes of wind direction and to collisions. Coloured contours extracted from GPS data are superimposed to aerial photographs taken from a plane or a kite (Supplementary Information). The averaged sand flux rose (1999, $Q = 80 \text{ m}^2 \text{ yr}^{-1}$) shows a narrow unimodal regime whose dominant direction, indicated by yellow arrows, is the average direction of barchan motion. The maximum sand flux rose reveals a more isotropic distribution of extreme events (Supplementary Information). **a–d**, Response of barchan dunes to changes of wind direction. **a**, Two profiles of a small dune (purple, $t = -15$ months; green, $t = -12$ months) superimposed on an aerial photograph ($t = 0$). The slip face quickly readapts to changes of wind direction (corresponding arrows). **b**, Surface waves on a medium-sized dune, induced by a northwest wind (reference profile at $t = -80$ months). **c**, Destabilization of a medium-sized dune leading to the emission of elementary barchans ($t = 50$ months).

d, Mega-barchan ($t = 350$ months). The surface waves are permanently driven by daily wind variations. **e**, Numerical dune produced by the C_C^C model⁸. The parameters ($A = 8.4$, $B = 8.0$, $D = 0.4$) are tuned to get the same height, length and width as on **c**. **f–h**, Response of barchan dunes to collisions. **f**, Edging collisions between small dunes (yellow, $t = 17$ months, green, $t = 23$ months, purple, $t = 33$ months). Note also a fusion (red arrow). **g**, Coaxial collision of medium-sized dunes leading to a larger dune with a wake of new elementary barchans ($t = 60$ months). **h**, Chain destabilization of medium-sized dunes ($t = 60$ months). The trail of a dune sustains the instability of the next one. **i**, Long-term ($t = 350$ months) evolution of a barchan field. Even though the largest dunes (highlighted dashed contours) persist, many dunes appear or disappear, showing that dune destabilization is an essential process of large-scale rearrangement.

framework of the C_C^C model, the linear stability analysis of a uniform sand bed predicts an unconditional instability at large wavelengths of growth rate σ and propagation speed c given by:

$$\sigma = Qk^2 \left[\frac{B - A|k|L}{1 + (kL)^2} - D \right] \quad c = \frac{Q|k|}{1 + (kL)^2} (A + B|k|L) \quad (2)$$

where $k = 2\pi/\lambda$ is the mode wavenumber. As all modes propagate downwind, the instability is of convective type¹⁶ so that the pattern develops spatially: any disturbance such as a change of wind direction or a colliding dune (Fig. 1) constitutes a seed amplified by the linear instability in the course of propagation.

The maximum growth rate—the most unstable mode predicted by equation (2)—is reached for a wavelength $\lambda_{\text{max}} \approx 12L \approx 20 \text{ m}$. λ was measured systematically in a zone of 500 regular barchans (1–12 m high). The histogram of these measurements, depicted on Fig. 3a, exhibits a clear peak around $\lambda \approx 20 \text{ m}$. The same measurements, performed on the windward side of the mega-barchan of Fig. 1d (40 m high), give a slightly larger wavelength of $\lambda \approx 28 \text{ m}$. This can be related to an increase of the wavelength—pattern coarsening—as the waves propagate, due to screenings and fusions. Situations where the fluid is much denser (water) or lighter (Mars

atmosphere) than air allows a further examination of the scaling proposed in equation (1). It gives the correct destabilization wavelength for submarine ripples in the turbulent case: λ_{max} reduces to a few centimetres, because water is 800 times denser than air. Martian dunes¹⁷ present surface waves of wavelength around 600 m for an atmosphere 80 times lighter than Earth's. According to our analysis, the grains in saltation forming these dunes should have a diameter of $75 \mu\text{m}$, which corresponds to the size of the dominant species composing the aeolian ripples photographed by the Mars rover Opportunity¹⁸.

As shown in Fig. 2a, the velocity of large barchans decreases with height H following the Bagnold relation¹ $c \approx aQ/H$, where $a \approx 2.7$ represents the increase in flux between the ground and the brink and is measured independently (Supplementary Information). Furthermore, we expect the velocity of barchans to match that of waves in the limit of vanishing amplitude, for which equation (2) predicts: $c \approx c(\lambda_{\text{max}}) \approx 5.3Q/L$. This suggests that we should modify the Bagnold relation:

$$c = aQ/(H + H_0) \quad (3)$$

where $H_0 = aL/5.3 \approx 87 \text{ cm}$ is a cut-off that corresponds to the

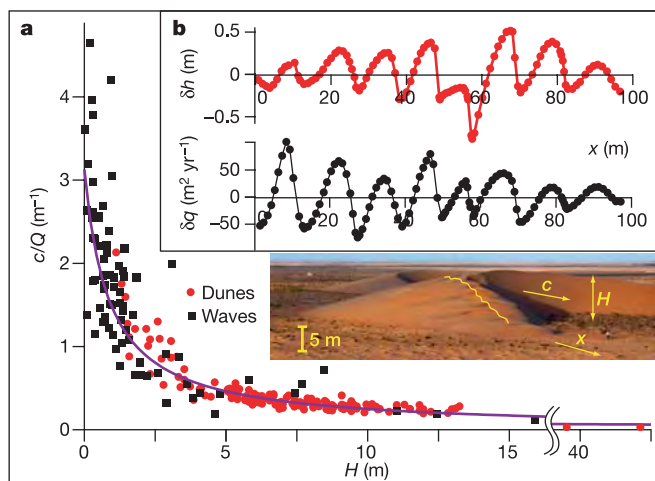


Figure 2 | Wave-like behaviour of surface undulations. **a**, Comparison of the velocity c of waves and dunes as a function of their height H . The velocity of 140 dunes is measured using several aerial photographs and global positioning system (GPS) data, with displacements resolved to within 4 m. The velocity is averaged over two time intervals (300 to 350 months and 10 to 18 months) to sample both large dunes that do not propagate much over a year, and small ones that cannot be followed over long periods of time (see Fig. 1i). The waves were marked with sticks and their displacements measured after a few hours. The corresponding velocities follow the same law as that of dunes, given a redefinition of Q : it is the saturated flux on solid ground averaged over the period of time considered in the case of dunes, and in the case of waves, the local saturated flux estimated from their altitude and from the propagation of neighbouring barchans (Supplementary Information). The purple solid line corresponds to the prediction of equation (3). **b**, Comparison of a wave profile δh with the corresponding modulations of sand flux δq . Sticks are placed regularly along a cut of a barchan horn (yellow line on the photograph). The slope and the distance between sticks allow for the reconstruction of the height profile δh . δq is derived from the erosion rate $\partial_t h$, using mass conservation: $\delta q = -\int \partial_t h dx$ (Supplementary Information). The oscillations of δq in phase with those of δh are characteristic of a plane propagating wave $h(x - ct)$ for which $\delta q = \int c \partial_t h dx = c \delta h$. The local wave velocity $c = \delta q / \delta h$ decreases down-slope, like the local flux Q .

minimal barchan height. Field data (Fig. 2a) shows a quantitative agreement with this model equation for the slow motion of barchans on solid ground, but also for the fast motion of waves on the surface of dunes. It demonstrates that large-amplitude waves behave as barchans.

As for any convective instability, the amplitude of the waves depends on the amplitude of the disturbances applied to the dune, as well as on the length of the dune itself. This is well illustrated by the response of dunes to changes of wind direction. The medium-sized dune of Fig. 1c can remain almost undisturbed because it is sufficiently isolated and the wind sufficiently constant. Whereas dunes of size $\lambda_{\max}$ are small enough to quickly readapt their shape (Fig. 1a), larger ones ($\sim 5\text{--}10\lambda_{\max}$) generate a certain number of waveforms in the lee flank (Fig. 1b, c). The mega-barchan of Fig. 1d is so large ($30\lambda_{\max}$) that the daily wind variations ($\sim 15^\circ$) render its surface permanently unstable. Barchans and mega-barchans thus belong to the same class of dunes, the corrugations resulting from a size effect. A colliding dune constitutes another type of perturbation. Figure 1g shows a typical example of coaxial collision between two large dunes ($3.5\lambda_{\max}$ and $6\lambda_{\max}$), one catching the other up from behind. Displaying behaviour that is far from soliton-like (solitons are propagative solitary waves which do not interact with each other), the impacting dune induces large-amplitude waves on the flanks of the target dune, which amplify in the course of their propagation.

Figure 1i shows that these waves not only decorate the dunes, but

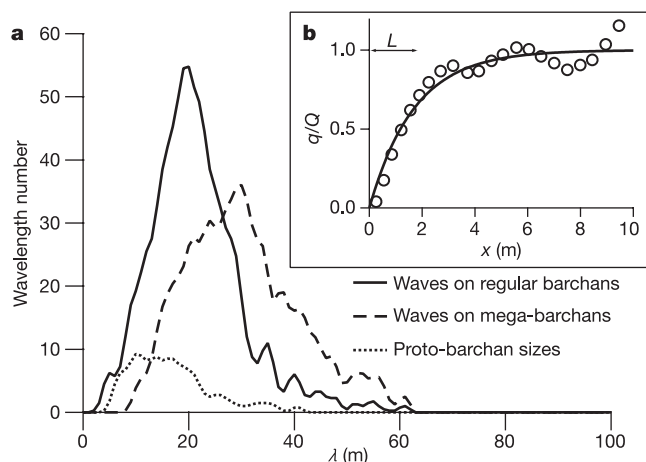


Figure 3 | Wavelength distribution and saturation length. **a**, Histograms of the wavelength λ measured systematically in a zone of $20\text{ km} \times 8\text{ km}$ (solid line, 930 data points), as well as on the windward side of the mega-dune shown on Fig. 1d (dashed line, 850 data points), using 1-m bins. The peak of the distribution is reached for $\lambda = 20$ and 28 m respectively, which is consistent with the linear stability analysis. The histogram of the length of proto-barchans—dome dunes without slip face—is also superimposed (dotted line, 160 data points). It shows a peak at 14 m . **b**, Spatial transient of sand transport saturation. A 20-m -long sand sheet is prepared with a bulldozer on the solid ground and protected to insure a null flux at the upwind edge. As for the wave profiles (Fig. 2b), the flux is measured by integrating the erosion rate $\partial_t h$, averaged over 24 h . The data are well fitted by an exponential relaxation $q = Q[1 - \exp(-x/L)]$, which corresponds to a first-order equation of saturation, $L \partial_x q = Q - q$.

play a crucial part in the long-term evolution of a barchan field. When a wave reaches solid ground at the end of the horn, an elementary dune of typical size slightly smaller than $\lambda_{\max}$ (Fig. 3a) is detached and emitted. Globally, this produces numerous newborn dunes in the wake of the collision and thus a large sand loss from the procreative dunes. There are only two circumstances under which this collision mechanism apparently looks like a non-interactive process (1) the case where the two interacting dunes are themselves of lengths comparable to the elementary size $\lambda_{\max}$ and (2) the case where they only brush against each other (Fig. 1f). We have demonstrated⁸ that dunes behaving as stable kinematic waves would not lead to the size selection observed in barchan fields. On the contrary, the whole dune field would merge into a single giant dune. To resolve this problem, there should exist a sand leak increasing with the dune size: our field study shows that the induction of waves on the surface of dunes, breaking the horns into elementary barchans, provides this missing mechanism.

Received 18 December 2004; accepted 30 June 2005.

1. Bagnold, R. A. *The Physics of Blown Sand and Desert Dunes* (Methuen, London, 1941).
2. Wiggs, G. F. S., Livingstone, I. & Warren, A. The role of streamline curvature in sand dune dynamics: evidence from field and wind tunnel measurements. *Geomorphology* **17**, 29–46 (1996).
3. Lancaster, N., Nickling, W. G., McKenna-Neuman, C. K. & Wyatt, V. E. Sediment flux and airflow on the stoss slope of a barchan dune. *Geomorphology* **17**, 55–62 (1996).
4. McKenna-Neuman, C., Lancaster, N. & Nickling, W. G. Relations between dune morphology, air flow, and sediment flux on reversing dunes, Silver Peak, Nevada. *Sedimentology* **44**, 1103–1113 (1997).
5. Finkel, H. J. The barchans of southern Peru. *J. Geol.* **67**, 614–647 (1959).
6. Hastenrath, S. L. The barchans of the Arequipa region, southern Peru. *Z. Geomorphol.* **11**, 300–331 (1967).
7. Andreotti, B., Claudin, P. & Douady, S. Selection of dune shapes and velocities—Part 1: Dynamics of sand, wind and barchans. *Eur. Phys. J. B* **28**, 321–339 (2002).
8. Hersen, P. et al. Corridors of barchan dunes: stability and size selection. *Phys. Rev. E* **69**, 011304 (2004).

9. Besler, H. Complex barchans in the Libyan Desert: dune traps or overtaking solitons? *Z. Geomorph. N.F. Suppl.-Bd.* **126**, 59–74 (2002).
10. Schwammle, V. & Herrmann, H. J. Solitary wave behaviour of sand dunes. *Nature* **426**, 619–620 (2003).
11. Katsuki, A., Nishimori, H., Endo, N. & Taniguchi, K. Collision dynamics of two barchan dunes simulated by a simple model. *J. Phys. Soc. Jpn* **74**, 538–541; preprint at (<http://arXiv.org/cond-mat/0403312>) (2005).
12. Livingstone, I., Wiggs, G. F. S. & Baddock, M. Barchan dunes: why they cannot be treated as solitons or solitary waves. *Earth Surf. Process. Landforms* **30**, 255–257 (2005).
13. Fryberger, S. G. & Dean, G. in *A Study of Global Sand Seas* (ed. McKee, E. D.) 137–169 (Geol. Surv. Prof. Pap. 1052, Washington, 1979).
14. Hersen, P., Douady, S. & Andreotti, B. Relevant length scale of barchan dunes. *Phys. Rev. Lett.* **89**, 264301 (2002).
15. Andreotti, B. A two-species model of aeolian sand transport. *J. Fluid Mech.* **510**, 47–70 (2004).
16. Huerre, P. & Monkewitz, P. P. Local and global instabilities in spatially developing flows. *Annu. Rev. Fluid Mech.* **22**, 473–537 (1990).
17. European Space Agency. *Mars Express* (http://www.esa.int/SPECIALS/Mars_Express/) (23 June 2005).
18. Mars Rover Blog. *Mars Rover Photos* (<http://www.markcarey.com/mars/mars-rover-photos.html>) (23 June 2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Bellot and L. Quartier for assistance during field measurements, R. Hoyle, B. Murray and J. Snoeijer for critical reading of the manuscript, and H. Tsoar for the provision of sand roses. This study was supported by an ACI Jeunes Chercheurs.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.A. (andreotti@pmmh.espci.fr).

LETTERS

Trace element signature of subduction-zone fluids, melts and supercritical liquids at 120–180 km depth

Ronit Kessel^{1,2}, Max W. Schmidt², Peter Ulmer² & Thomas Pettke³

Fluids and melts liberated from subducting oceanic crust recycle lithophile elements back into the mantle wedge, facilitate melting and ultimately lead to prolific subduction-zone arc volcanism^{1,2}. The nature and composition of the mobile phases generated in the subducting slab at high pressures have, however, remained largely unknown^{3–7}. Here we report direct LA-ICPMS measurements of the composition of fluids and melts equilibrated with a basaltic eclogite at pressures equivalent to depths in the Earth of 120–180 km and temperatures of 700–1,200 °C. The resultant liquid/mineral partition coefficients constrain the recycling rates of key elements. The dichotomy of dehydration versus melting at 120 km depth is expressed through contrasting behaviour of many trace elements (U/Th, Sr, Ba, Be and the light rare-earth elements). At pressures equivalent to 180 km depth, however, a supercritical liquid with melt-like solubilities for the investigated trace elements is observed, even at low temperatures. This mobilizes most of the key trace elements (except the heavy rare-earth elements, Y and Sc) and thus limits fluid-phase transfer of geochemical signatures in subduction zones to pressures less than 6 GPa.

Trace elements and isotope ratios of arc magmas are frequently deployed to decipher their complex magmatic history and identify source components. It is well established that a slab component, identified by B, ¹⁰Be, U and Th abundances^{5–10}, combines with a mantle melting component with high Mg# and high Ni and Cr contents¹¹, to form primitive basaltic magmas that give rise to most arc suites. Whether at high pressures the slab component, which transfers such elements to the mantle, is a fluid, melt or supercritical liquid, has remained unknown. Most thermal models of mature subduction zones¹² suggest that temperature conditions for the subducting crust are far below its solidus at subarc depths. Nevertheless, it has become increasingly popular to attribute estimated Be and Th transfer rates to slab melting⁴, implying that most thermal models are in error. The missing link in this controversy is the composition and nature of the fluids or melts departing from the slab, which in turn allow conditions of their formation to be constrained. A critical petrological parameter in this debate is the closure of the immiscibility gap between an aqueous fluid and a hydrous silicate melt^{13,14}.

At typical crustal pressures, aqueous fluids contain a few per cent of solute while partial melts have <15 wt% H₂O. At higher pressures, the amount of solutes in the fluid and the solubility of H₂O in silicate melts both dramatically increase, and the compositions of fluid and melt converge along a miscibility gap, which eventually disappears¹⁵. The intersection of the miscibility gap's critical curve with the solidus then defines the end point of the solidus^{14–16}, and we term any fluid or melt beyond this point 'supercritical liquid'. The pressure of the critical end point of the solidus is strongly dependent on bulk

composition, and varies from 1.6 GPa in feldspathic systems¹⁶ to >10 GPa in mantle systems¹⁵. In potassium- and mica-rich lithologies of the subducted crust, this endpoint is situated between 5 and 6 GPa (ref. 14), corresponding to depths of 150–180 km of the subducting slab beneath volcanic arcs. It has been proposed³ that supercritical liquids have high trace element solubilities at relatively low temperatures. The scope of our experimental study is to quantify partitioning between the residual slab mineralogy and all three mobile phases of interest—that is, high pressure aqueous fluids, hydrous melts and supercritical liquids.

To determine fluid and melt compositions equilibrated with an eclogitic residue, we have employed two novel techniques. We performed standard diamond-trap experiments^{17,18} (see Methods) on an average mid-ocean-ridge basalt (MORB), but used a 'rocking' multi-anvil apparatus¹⁹. In this device, the entire multi-anvil is turned upside-down regularly, such that the fluid in the capsule is constantly homogenized by induced convection. The turning has been shown to be critical for maintaining a homogeneous fluid¹⁹ over the entire duration of the experiment (12–72 h), thus maintaining equilibrium between the fluid and all minerals. The fluid-saturated K-free MORB was doped with 26 trace elements, and equilibrated at pressures of 4–6 GPa and temperatures of 700–1,200 °C. After quenching of the experiments, the recovered capsule was placed in a freezing device. The frozen capsule was then cut open, and transferred to the laser-ablation cell for immediate laser-ablation inductive-coupled-plasma mass spectrometry (LA-ICPMS) analysis of the ice in the diamond trap, keeping it frozen for the whole duration of the analysis. This technique overcomes the hitherto unavoidable loss of some solute and precipitates upon exsolution of the high pressure fluids/liquids/melts when the fluid is released from the capsule (see Methods). Thus, for the first time, we accurately determine major and trace element concentrations and H₂O contents of high pressure fluids and unquenchable melts²⁰. After analysis of the trapped ice, capsules were defrosted, mounted, and polished sections of each run were examined using scanning electron microscopy. The major element compositions of the eclogitic minerals were determined by electron microprobe, and trace element concentrations by conventional LA-ICPMS.

Combining these two methods, we have determined compositions of the mobile phases and partition coefficients between the mobile phase and eclogite minerals, thus facilitating modelling of element mobilities in the subducting slab. The experimental results show that at 4 GPa, fluid saturated melting takes place, in which a fluid that increases its solute content from 13 to 22 wt% between 700 and 900 °C (Table 1) is replaced by a silicate melt with ≤56 wt% H₂O at ≥1,000 °C (ref. 20). At 5 GPa, this transition occurs at 1,000–1,050 °C, but at 6 GPa, the behaviour is different. A gradual decrease in H₂O concentrations in the mobile phase from 72 to 31 wt% H₂O

¹Institute of Earth Science, The Hebrew University of Jerusalem, Jerusalem, 91904, Israel. ²Institute für Mineralogie und Petrographie, ³Isotope Geochemistry and Mineral Resources, ETH Zentrum, Zürich, CH-8092, Switzerland.

Table 1 | Bulk average MORB (starting material), fluid, melt and supercritical liquid compositions

Conc. (wt%)	Bulk	Pressure/temperature conditions*			
		4/800	4/1,000	6/800	6/1,000
SiO ₂	52.84	12.0(1.2)‡	31.4(3.5)	19.8(6.5)	29.8(6.6)
TiO ₂	1.43	0.04(1)	0.50(3)	0.11(3)	0.71(16)
Al ₂ O ₃	17.05	1.45(49)	4.24(30)	1.41(80)	2.89(68)
FeO†	8.41	0.22(8)	1.27(21)	0.63(49)	1.12(30)
MgO	5.87	0.56(27)	0.78(6)	0.68(36)	0.85(21)
CaO	10.11	0.84(45)	2.05(12)	1.93(88)	3.25(74)
Na ₂ O	3.20	0.96(25)	2.80(35)	1.46(60)	2.64(1.62)
H ₂ O	—	83.1(2.5)	56.1(3.8)	72.6(8.7)	57.6(9.3)

* x/y: experiment at pressure x (in GPa) and temperature y (in °C).
† All Fe is presented as FeO.
‡ Numbers enclosed in parentheses indicate the absolute value of one standard deviation of the distribution of average compositions reported in terms of the least digits cited; that is, 12.0(1.2) indicates 12.0 ± 1.2.

between 800 and 1,300 °C is observed²⁰. These results demonstrate a chemical continuum between fluid and melt at 6 GPa, and hence conditions beyond the critical endpoint of the solidus. The supercritical liquid replaces the classic dichotomy of fluid and melt at lower pressures.

Our results on the partitioning of trace elements (Fig. 1) show a distinct change in pattern when moving from fluid and melt at 4 GPa to a supercritical liquid at 6 GPa. A fluid at 4 GPa still behaves similarly to fluids from lower pressures^{4,21,22}, but with most bulk partition coefficients $D^{\text{fluid/solid}}$ increased, thus increasing the mobility for most elements. Aqueous fluids resulting from dehydration of an eclogite at 4 GPa/700–800 °C are characterized by $D^{\text{fluid/solid}}$ of Rb, Cs, Ba and Pb of 10 or higher, by $D^{\text{fluid/solid}}_{\text{U}} \geq D^{\text{fluid/solid}}_{\text{Th}}$ (Fig. 2), by $D^{\text{fluid/solid}}_{\text{REE}}$ and $D^{\text{fluid/solid}}_{\text{HFSE}}$ between 0.01 and 0.5, and a limited mobility of Li (here REE indicates rare-earth elements, and HFSE high field strength elements). In contrast, the hydrous melt at 4 GPa/1,000 °C has $D^{\text{melt/solid}}_{\text{U}} < D^{\text{melt/solid}}_{\text{Th}}$ and a steep REE pattern characteristic for garnet in the residue. There is no principal difference in the partitioning trends of these elements from partial melts of eclogite at lower

pressure^{4,23,24}. At 4 GPa, the partition coefficients $D^{\text{(fluid or melt)/solid}}$ of three key elements—Be, Sr and Th—display a very strong temperature dependence, but without a conspicuous jump at the solidus (Fig. 1a). The mobility of these elements (see Supplementary Tables) in the experiments is <15% at 700 °C and gradually increases to 60–90% at 1,000 °C.

At 6 GPa, supercritical liquids are characterized by a high mobility of almost all of the investigated elements except heavy rare-earth elements (HREE), Y and Sc, whose $D^{\text{liquid/solid}}$ remain in the range 0.01–0.5 (Fig. 1b). In the supercritical liquids, $D^{\text{liquid/solid}}_{\text{U}}$ is $< D^{\text{liquid/solid}}_{\text{Th}}$, and $D^{\text{liquid/solid}}_{\text{Be}}$ ranges from 6 to 22. In conjunction with the extreme fractionation of REE (for example, $D^{\text{liquid/solid}}_{\text{La}}/D^{\text{liquid/solid}}_{\text{Lu}} \approx 10^3$) and the highly mobile Sr (Figs 1b and 2a), a trace element pattern similar to lower pressure silicate melts emerges, although this pattern is produced by a supercritical liquid with 72–56 wt% H₂O (ref. 20) over the entire investigated temperature range (800–1,200 °C). In particular, the mobility of Be increases dramatically with pressure. Although fairly insoluble at 2 GPa/900 °C ($D^{\text{fluid/eclogite solid}}_{\text{Be}} < 0.5$; ref. 25), $D^{\text{fluid/solid}}_{\text{Be}}$ increases to 1–3 at 4 GPa/800–900 °C, and to 6–8 at 6 GPa/800–900 °C, leading to >60% of the Be being dissolved in the supercritical liquid of our experiments at temperatures of 800 °C. Fluids are clearly distinguished from melts and supercritical liquids (for example, U/Th, Th/Rb, La/Lu), but none of the ratios of the investigated trace elements can be used unequivocally to discriminate between trace element transfer through hydrous melts or supercritical liquids. The supercritical liquids exhibit in fact an ‘adakite-type pattern’²⁶ with high LREE/HREE, Sr/HREE and Sr/Y, and with excess Th (ref. 8) through a petrological process different from the melting of metabasaltic or metasedimentary eclogites at moderate pressures. This adds further evidence that high bulk Mg#, Ni and Cr for a given SiO₂ content of an arc lava are the only reliable indicators of andesitic to dacitic melts passing through the mantle wedge³.

In our experiments, the solubilities of Nb, Ta, Zr and Hf in garnet decrease strongly with temperature. While garnet retains >70% of Nb, Ta, Zr and Hf at 700–800 °C, these elements become

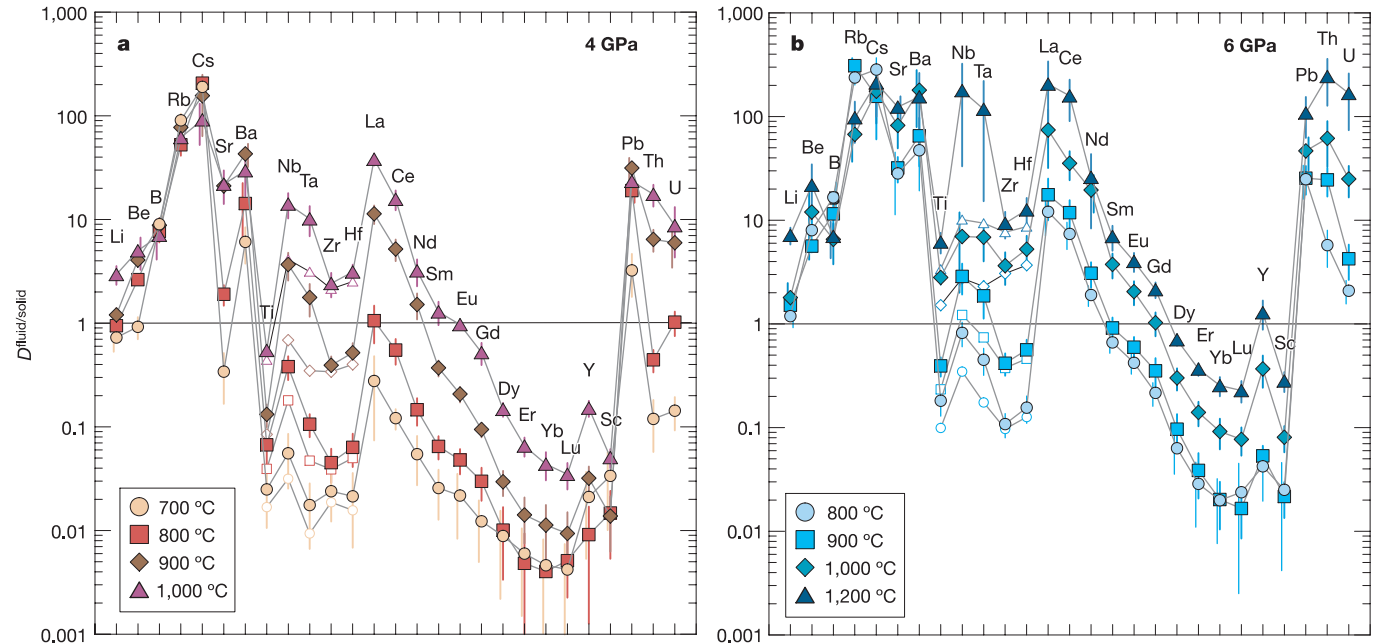


Figure 1 | Experimental fluid-solid partition coefficients for average MORB. a, Aqueous fluids (700–900 °C) and hydrous melt (1,000 °C) at 4 GPa; **b**, supercritical liquid (800–1,200 °C) at 6 GPa. The solid residue is calculated from the element concentrations in the solids multiplied by their abundance in the residue. Filled symbols refer to a garnet-clinopyroxene

residue, open symbols to a garnet-clinopyroxene-rutile residue (see also Supplementary Information). Uncertainties are given as 1σ and were calculated by propagating uncertainties of elemental concentrations. 1σ error bars are sometimes shifted off the symbol centre for clarity.

incompatible and enter the silicate melt or supercritical liquid at $\geq 1,000^\circ\text{C}$. However, as at lower pressures, Nb, Ta and Ti may be held back by residual rutile²¹. In the investigated pressure range, TiO_2 solubilities in the liquid are mostly temperature dependent, and significant amounts of rutile (for average MORB decreasing from 0.5 wt% at 900°C to respectively 0.3 and 0.2 wt% at 1,000 and $1,200^\circ\text{C}$) will produce Ti, Nb and Ta anomalies typical for most arc magmas. We suggest that the extent of this anomaly, combined with the abundances of Be and Th in primary arc magmas, are probably the best constraints on the temperature during release of the mobile slab component.

Beryllium, Th, U, and radiogenic Pb and Sr are enriched in pelitic sediments and greywackes⁴, thus dominating the budget and transfer rates of these elements. These sediments have the same eclogitic mineralogy (but different mineral proportions) as basaltic rocks at depths $>100\text{ km}$ (ref. 14): in particular, they have up to 25 wt%

phengite owing to much higher bulk K_2O contents. The high solubility of K_2O in hydrous liquids has the result that a few per cent of H_2O entirely dissociate phengite¹⁴, with all K_2O dissolved in the mobile phase. Thus, any significant mass transfer process will remove phengite from the residue and our data can thus be equally applied to pelitic and clastic metasediments. Fluid/rock ratios in the metasediments located on top of the subducting lithosphere are intrinsically high, as any fluid liberated below must pass through the comparatively thin sediment pile. Although fewer H_2O sources are available in hydrated metabasites at 4–6 GPa than at $\leq 3\text{ GPa}$ (ref. 27), a major H_2O reservoir at these depths is the partially hydrated peridotite underlying the oceanic crust. Serpentine and chlorite²⁸ therein may reach their stability limits, dehydrate, and flush the oceanic crust. Geochemically effective H_2O /sediment ratios are a result of the relative masses of the major H_2O reservoirs (MORB, serpentinite) and the fluid flow mechanism (pervasive versus channelled flow). Both dehydration and flushing are incremental processes significantly increasing mobilities for all elements with $D^{\text{fluid/solid}}$ greater than unity. It is thus safe to conclude that a large fraction of Be, B, Rb, Cs, Sr, Ba, LREE, Pb, Th and U will be removed effectively from the sediments by supercritical liquids.

Further complications in the production of a mobile phase may arise from carbonate saturated sedimentary or altered mafic crust compositions^{29,30}. Characteristically, at high pressure/low temperature conditions, these have dolomite or magnesite in addition to the eclogitic garnet + clinopyroxene $\pm$ coesite $\pm$ kyanite $\pm$ phengite assembly^{30,31}. Most important, the only significant carbonaceous species at these conditions is CO_2 , and the molar fraction of CO_2 in a carbonate saturated aqueous fluid is low (≤ 0.10) at subsolidus conditions²⁹, resulting in a minor shift of the fluid saturated solidus of carbonated basaltic eclogite³¹ of less than 50°C . Likewise, the addition of carbonates is expected to have a small effect on the pressure of the critical endpoint of the solidus and on trace element partition coefficients in the fluid or supercritical liquid.

The transport properties of supercritical liquids at moderate temperatures result in high Be and Th transfer rates similar to those observed in many arcs⁵ that so far have been attributed to melting of the sediment layer of the subducting crust⁴. Similarly, excess Th over U was thought to be a unique characteristic of slab melting. We have demonstrated here that such melt-like characteristics can be produced by supercritical high-pressure liquids at any temperature. We argue that the slab component in these arc magmas may result from supercritical liquids rather than from relatively shallow melts from a sediment source, thus relaxing the apparent temperature discrepancy between geochemical arguments and geophysical models. Finally, the chemical properties of supercritical liquids liberated at depths exceeding $\sim 160\text{ km}$ removes the dichotomy of fluids versus melts. As a consequence, fluid-type geochemical signatures from the oceanic crust are limited to depths $\leq 160\text{ km}$: at higher pressures, the mobile phase will inevitably add a melt-like trace element pattern to the mantle wedge.

METHODS

Experimental. A K-free average MORB composition³² was synthesized as a glass at an oxygen fugacity close to fayalite–magnetite–quartz (FMQ) at the experimental conditions. The glass was doped with 860 p.p.m. Cs, which is necessary as an internal standard for LA-ICPMS analyses. 26 trace elements were fused into a second glass of diopside composition, and about 2% of the diopside glass was mixed with the basalt powder such that individual bulk trace element concentrations were 100–250 p.p.m. Water (17–26 wt% H_2O) was first pipetted into the 2.3 mm Au capsule followed by $\sim 3\text{--}5\text{ mg}$ basalt powder, overlain by a thin layer of synthetic diamond aggregates (15–25 μm diameter grains), and another layer of doped basalt powder. Capsules were welded shut while cooled with liquid nitrogen to avoid H_2O loss.

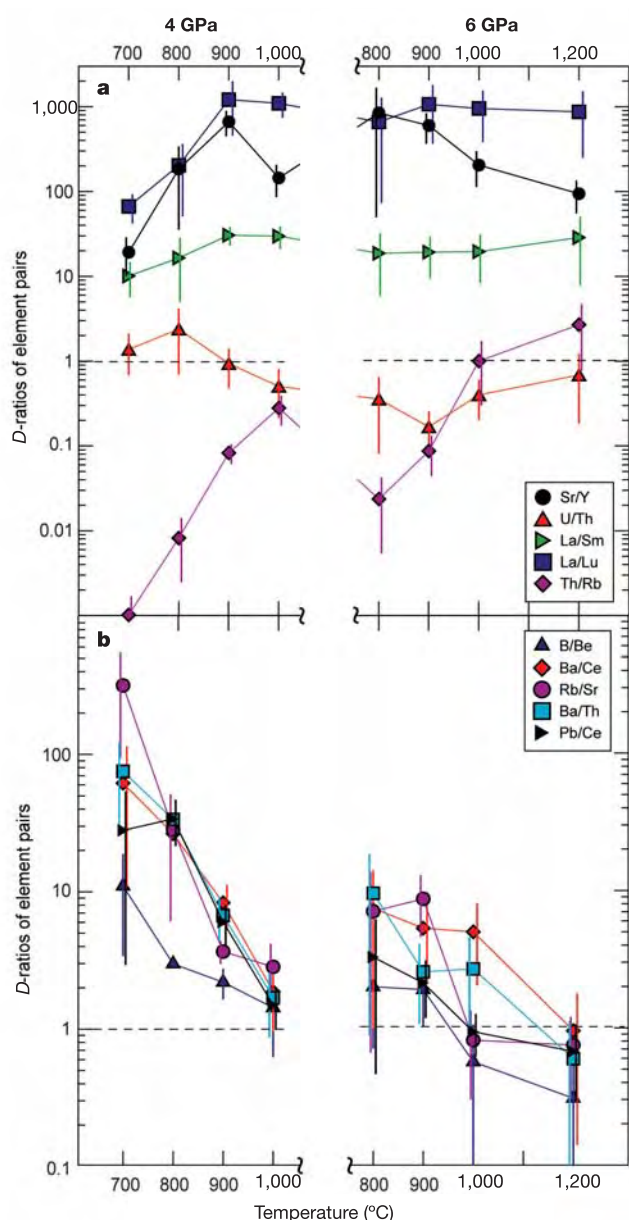


Figure 2 | Key trace element distribution coefficients characterizing the geochemical signature of the mobile phase during subduction processes. Uncertainties are given as 1σ and were calculated by propagating uncertainties of elemental concentrations. 1σ error bars are sometimes shifted off the symbol centre for clarity.

Experiments were conducted in a Walker-type multi-anvil apparatus employing a 19 mm MgO octahedron and WC cubes with a 12 mm truncation (each experiment used a different MgO octahedron). The Au capsule was placed in a BN cylinder inside a stepped graphite furnace, filled with MgO spacers and Mo caps. To avoid compositional zoning, the multi-anvil apparatus was mounted on a rocking device enabling the apparatus to be turned 180° while maintaining pressure and temperature¹⁹. Frequent inversion (every 15 min) of the charge in the gravitational field leads to convection in the fluid and thus to re-homogenization and migration of the fluid phase in the capsule. The re-homogenization of the fluid counteracts Soret diffusion, which otherwise would automatically lead to zonation and disequilibrium between the fluid and one or several of the solids which would segregate into growth zones. This method allows for chemical equilibrium between all solids and the fluid and as a side effect enhances reaction rates.

Analytical. After the experiment, we analyse the diamond trap and the minerals as described above. The decisive advantage of the freezing method³² is that it allows us to determine directly the composition of unquenchable liquids. Such liquids constitute a mud of quench phases and water that maintains considerable amounts of trace elements (for example, of Cs, Rb, K) in solution. The traditional method of piercing the capsule releases a milky fluid that contains all of the solute and part of the quench crystals. Any further preparation step results in additional losses of quench material. We thus avoid all of these steps by maintaining the fluid in the capsule in its solid state during the LA-ICPMS analyses, to allow ablation of the complete fraction of the then unmixed liquid. The largest suitable ablation spot (60–110 µm diameter) is used to analyse the diamond trap in order to obtain the best averaging during sampling of the experimental product interstitial to the ~20 µm diameter diamonds. A highly incompatible element (Cs), fractionating completely into the fluid or melt phase, is used as an internal standard for LA-ICPMS signal; at a known Cs:H ratio of the starting material, we can quantify the H₂O content of the unquenchable liquid that has within the error the same Cs:H ratio, which in turn enables us to determine absolute concentrations of major and trace elements in the unquenchable liquid. Details of the LA-ICPMS analytics are found in ref. 32.

Received 3 February; accepted 24 June 2005.

- Kushiro, K. in *Magmatic Processes: Physicochemical Principles* (ed. Mysen, B. O.) 165–181 (Special Publication No. 1, Geochemical Society, 1987).
- Tatsumi, Y., Hamilton, D. L. & Nesbitt, R. W. Chemical characteristics of fluid phase released from a subducted lithosphere and origin of arc magmas: evidence from high-pressure experiments and natural rocks. *J. Volcanol. Geotherm. Res.* **29**, 293–309 (1986).
- Kelemen, P. B., Hanghj, K. & Greene, A. R. One view of the geochemistry of subduction-related magmatic arcs, with an emphasis on primitive andesite and lower crust. *Treatise Geochem.* **3**, 593–659 (2004).
- Johnson, M. C. & Plank, T. Dehydration and melting experiments constrain the fate of subducted sediments. *Geochim. Geophys. Geosyst.* **1**, doi:10.1029/1999GC000014 (1999).
- Hawkesworth, C. J., Turner, S. P., McDermott, F., Peate, D. W. & van Clasteren, P. U-Th isotopes in arc magmas: Implications for element transfer from the subducted crust. *Science* **276**, 551–555 (1997).
- Morris, J. D., Leeman, W. P. & Tera, F. The subducted component in island arc lavas: constraints from Be isotopes and B-Be systematics. *Nature* **344**, 31–36 (1990).
- Elliot, T., Plank, T., Zindler, A., White, W. & Bourdon, B. Element transport from slab to volcanic front at the Mariana arc. *J. Geophys. Res.* **102**, 14991–15019 (1997).
- Sigmarsson, O., Martin, H. & Knowles, J. Melting of a subducting oceanic crust from U-Th disequilibria in austral Andean lavas. *Nature* **394**, 566–569 (1998).
- Ryan, J. G., Morris, J., Tera, F., Leeman, W. P. & Tsvetkov, A. Cross-arc geochemical variations in the Kurile arc as a function of slab depth. *Science* **270**, 625–627 (1995).
- Morris, J. D. & Ryan, J. G. Subduction zone processes and implications for changing composition of the upper and lower mantle. *Treatise Geochem.* **2**, 451–470 (2004).
- Tatsumi, Y. & Eggins, S. *Subduction Zone Magmatism* (Frontiers in Earth Sciences, Blackwell Science, Cambridge, Massachusetts, 1995).
- Kincaid, C. & Sachs, I. S. Thermal and dynamical evolution of the upper mantle in subduction zones. *J. Geophys. Res.* **102**, 12295–12315 (1997).
- Turner, S. & Hawkesworth, C. Constraints on flux rates and mantle dynamics beneath island arcs from Tonga-Kermadec lava geochemistry. *Nature* **389**, 568–573 (1997).
- Schmidt, M. W., Vielzeuf, D. & Auzanneau, E. Melting and dissolution of subducting crust at high pressures: the key role of white mica. *Earth Planet. Sci. Lett.* **228**, 65–84 (2004).
- Stalder, R., Ulmer, P., Thompson, A. B. & Günter, D. High pressure fluids in the system MgO-SiO₂-H₂O under upper mantle conditions. *Contrib. Mineral. Petrol.* **140**, 607–618 (2001).
- Shen, A. H. & Keppler, H. Direct observation of complete miscibility in the albite-H₂O system. *Nature* **385**, 710–712 (1997).
- Baker, M. B. & Stolper, E. M. Determining the composition of high-pressure mantle melts using diamond aggregates. *Geochim. Cosmochim. Acta* **58**, 2811–2827 (1994).
- Kushiro, I. & Hirose, K. Experimental determination of composition of melt formed by equilibrium partial melting of peridotite at high pressures using aggregates of diamond grains. *Proc. Jpn. Acad.* **68**, 63–68 (1992).
- Schmidt, M. W. & Ulmer, P. A rocking multianvil: Elimination of chemical segregation in fluid-saturated high-pressure experiments. *Geochim. Cosmochim. Acta* **68**, 1889–1899 (2004).
- Kessel, R., Ulmer, P., Pettker, T., Schmidt, M. W. & Thompson, A. B. Experimental determination of phase relations and second critical endpoint in K-free eclogite-H₂O at 4–6 GPa and 700–1400 °C. *Earth Planet. Sci. Lett.* (in the press).
- Brenan, J. M., Shaw, H. F., Ryerson, F. J. & Phinney, D. L. Mineral-aqueous fluid partitioning of trace elements at 900 °C and 2.0 GPa: Constraints on the trace element chemistry of mantle and deep crustal fluids. *Geochim. Cosmochim. Acta* **59**, 3331–3350 (1995).
- Stalder, R., Foley, S. F., Brey, G. P. & Horn, I. Mineral-aqueous fluid partitioning of trace elements at 900–1200 °C and 3.0–5.7 GPa: New experimental data for garnet, clinopyroxene, and rutile, and implications for mantle metasomatism. *Geochim. Cosmochim. Acta* **62**, 1781–1801 (1998).
- Green, T. H., Blundy, J. D., Adam, J. & Yaxley, G. M. SIMS determination of trace element partition coefficients between garnet, clinopyroxene and hydrous basaltic liquids at 2–7.5 GPa and 1080–1200 °C. *Lithos* **53**, 165–187 (2000).
- Klemme, S., Blundy, J. D. & Wood, B. J. Experimental constraints on major and trace element partitioning during partial melting of eclogite. *Geochim. Cosmochim. Acta* **66**, 3109–3123 (2002).
- Brenan, J. M., Ryerson, F. J. & Shaw, H. F. The role of aqueous fluids in the slab-to-mantle transfer of boron, beryllium, and lithium during subduction: Experiments and models. *Geochim. Cosmochim. Acta* **62**, 3337–3347 (1998).
- Martin, H. Adakitic magmas: modern analogues of Archean granitoids. *Lithos* **46**, 411–429 (1999).
- Schmidt, M. W. & Poli, S. Experimentally based water budgets for dehydrating slabs and consequences for arc magma generation. *Earth Planet. Sci. Lett.* **163**, 361–379 (1998).
- Fumagalli, P. & Poli, S. Experimentally determined phase relations in hydrous peridotites to 6.5 GPa and their consequences on the dynamics of subduction zones. *J. Petrol.* **46**, 555–578 (2005).
- Kerrick, D. M. & Connolly, J. A. D. Metamorphic devolatilization of subducted marine sediments and the transport of volatiles into the Earth's mantle. *Nature* **411**, 293–296 (2001).
- Holloway, J. R. & Domanik, K. J. Experimental synthesis and phase relations of phengitic muscovite from 6.5 to 11 GPa in a calcareous metapelite from the Dabie Mountains, China. *Lithos* **52**, 51–77 (2000).
- Yaxley, G. M. & Green, D. H. Experimental demonstration of refractory carbonate-bearing eclogite and siliceous melt in the subduction regime. *Earth Planet. Sci. Lett.* **128**, 313–325 (1994).
- Kessel, R., Ulmer, P., Pettker, T., Schmidt, M. W. & Thompson, A. B. A novel approach to determine high-pressure high-temperature fluid and melt compositions using diamond-trap experiments. *Am. Mineral.* **89**, 1078–1086 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.K. (kessel@vms.huji.ac.il).

LETTERS

Diatom carbon export enhanced by silicate upwelling in the northeast Atlantic

John T. Allen^{1,2}, Louise Brown^{1,3}, Richard Sanders¹, C. Mark Moore¹, Alexander Mustard¹, Sophie Fielding¹, Mike Lucas¹, Michel Rixen⁴, Graham Savidge⁵, Stephanie Henson¹ & Dan Mayor¹

Diatoms are unicellular or chain-forming phytoplankton that use silicon (Si) in cell wall construction. Their survival during periods of apparent nutrient exhaustion enhances carbon sequestration in frontal regions of the northern North Atlantic. These regions may therefore have a more important role in the 'biological pump' than they have previously been attributed¹, but how this is achieved is unknown. Diatom growth depends on silicate availability, in addition to nitrate and phosphate^{2,3}, but northern Atlantic waters are richer in nitrate than silicate⁴. Following the spring stratification, diatoms are the first phytoplankton to bloom^{2,5}. Once silicate is exhausted, diatom blooms subside in a major export event^{6,7}. Here we show that, with nitrate still available for new production, the diatom bloom is prolonged where there is a periodic supply of new silicate: specifically, diatoms thrive by 'mining' deep-water silicate brought to the surface by an unstable ocean front. The mechanism we present here is not limited to silicate fertilization; similar mechanisms could support nitrate-, phosphate- or iron-limited frontal regions in oceans elsewhere.

Diatoms initially dominate the spring bloom in the northern North Atlantic^{8,9,10,11}. However, diatom production is usually short-lived⁹, limited by the supply of dissolved silicate (orthosilicic acid, Si(OH)₄) in the sun-lit (euphotic) zone^{12,13}. Once Si(OH)₄ is exhausted, aggregations of diatoms sink out of the euphotic zone, efficiently exporting CO₂ from the atmosphere into the ocean interior^{6,7,12}. But, if physical processes prolong the Si(OH)₄ supply, diatoms can continue to thrive, increasing the potential for new carbon export both to the deep ocean and to higher trophic levels^{14,15,16}.

In the northeast Atlantic Ocean, maximal diatom growth occurs at the start of the spring bloom when Si(OH)₄ concentrations are 6–8 mmol m⁻³. Concentrations of <2 mmol m⁻³ begin to limit production⁸ and control diatom blooms; by the summer, surface Si(OH)₄ is depleted¹⁷ to <1 mmol m⁻³.

The Iceland–Faeroes Front (IFF) is the longest boundary between northeastern Atlantic waters and arctic waters of the Nordic seas¹⁸. The steep density gradient associated with the change in water properties across the IFF is unstable, creating dynamic meanders to the front, and detached eddies¹⁸. Such complex upper-ocean flows drive 'biological patchiness' in the spring bloom¹⁹.

In May and June 2001, biological and physical data were collected in the vicinity of the IFF during two legs of the RRS *Discovery* cruise for the Faeroes–Iceland–Scotland Hydrographic and Environmental Survey (FISHES)²⁰. An image of sea surface chlorophyll taken by the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) on 23 May 2001 (Fig. 1) showed a marked bloom along the meanders of the IFF. *In situ* fluorometers confirmed that high concentrations of chlorophyll (5–7.5 mg m⁻³) were restricted to the frontal jet (Fig. 2).

Taxonomic analysis showed that the frontal bloom was composed of diatoms, principally *Chaetoceros* spp., *Thalassiosira* spp. and *Asterionellopsis glacialis* (refs 20, 21), whereas phyto-flagellates dominated to the north and south. Stable ¹⁵N-uptake rate measurements²⁰ showed a predominance of new production within the IFF ($f = \rho\text{NO}_3/\rho\text{N}_{\text{total}} = 0.56$) relative to the north (0.31) or south (0.30) of the front, where lower f -ratios were expected based on the prevalence of smaller phytoplankton.

Surface Si(OH)₄ concentrations were depleted everywhere^{21,22} except in the remnant deep winter mixed layers south of the IFF (Fig. 2). Surface nitrate concentrations were also considerably reduced in the frontal jet and in the stratified surface waters north of the front.

Two incubation experiments, measuring radioisotope tracer (³²Si(OH)₄) uptake (see Methods), were carried out in the IFF region²². At station N (25 May 2001), north of the IFF (Fig. 1), phytoplankton samples (Table 1) were principally comprised of phyto-flagellates, with diatoms (dominated by *Fragilariopsis oceanica*; ref. 21)

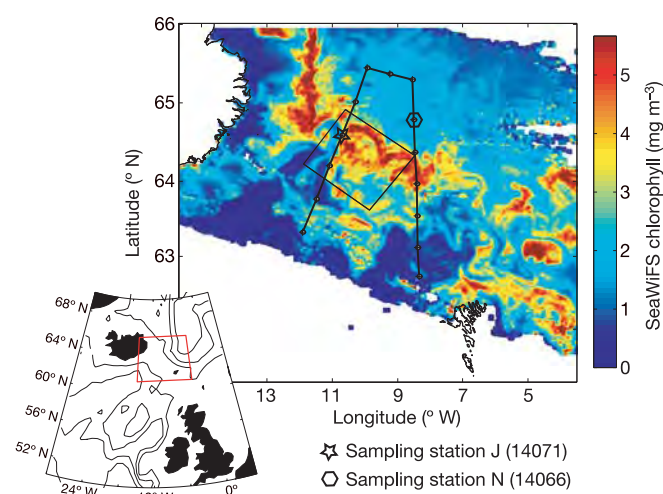


Figure 1 | Ocean colour-image of the northeast Atlantic. The inset map shows the general location of the study region superimposed on bathymetry of the area. The black lines on the main image show the relevant passage and stations around the IFF during leg one of the cruise, and a rectangle indicates the location of three repeated high-resolution SeaSoar³⁰ surveys during leg two of the cruise. These are superimposed on satellite-derived (SeaWiFS) sea surface chlorophyll concentrations for the region for 23 May 2001 (white areas are cloud and land). A marked phytoplankton bloom (yellow/orange/red) is associated with the front.

¹National Oceanography Centre, Southampton SO14 3ZH, UK. ²Environmental Research Institute, North Highland College, UHI Millennium Institute, Thurso, Caithness KW14 7JD, UK. ³School of Geography and Geosciences, University of St Andrews, Irvine Building, North Street, St Andrews, Fife KY16 9AJ, UK. ⁴NATO Undersea Research Centre, 19138 La Spezia, Italy. ⁵Queen's University of Belfast Marine Laboratory, Portaferry BT22 1PF, Northern Ireland, UK.

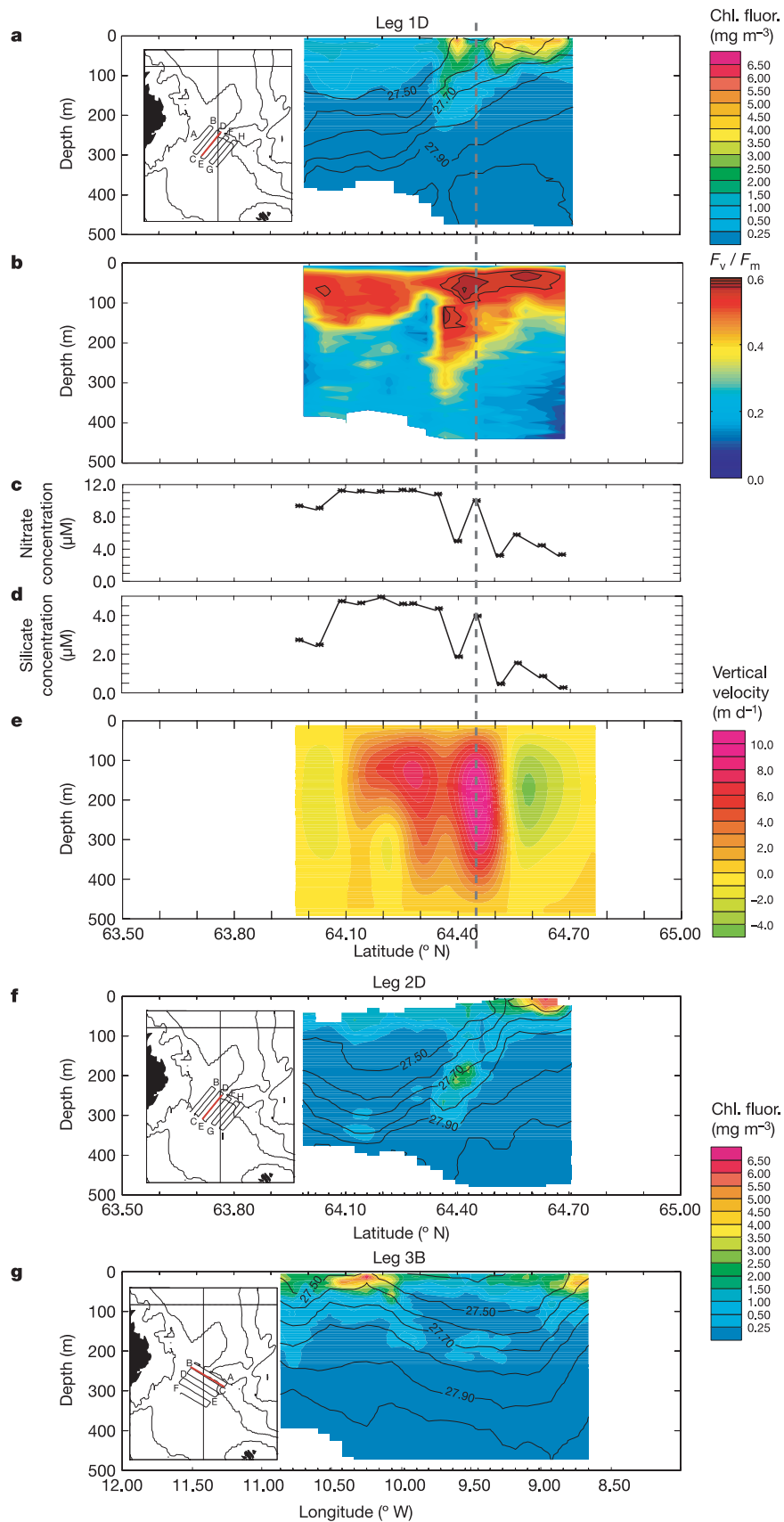


Figure 2 | Sections from each of the three SeaSoar surveys (7–17 June 2001). **a–e**, Chlorophyll fluorescence (Chl. fluor.; **a**), photosynthetic efficiency (F_v/F_m ; **b**), surface nitrate (**c**), surface silicate (**d**) and diagnosed vertical velocities (**e**) are compared for a leg of the first SeaSoar³⁰ survey. The highest three contour levels (0.58–0.60 and >0.60) for photosynthetic efficiency (F_v/F_m) are outlined for clarity. The grey dotted line indicates the approximate position of the front and highlights a strong correlation between upward vertical motion and a transient pool of upwelled $\text{Si}(\text{OH})_4$.

Photosynthetic efficiency was high (~ 0.55 – 0.60) within the diatom-dominated frontal bloom and lower in the flagellate-dominated communities to the north (0.4 – 0.5 , data not shown²¹) and south (~ 0.5) of the front. **f**, **g**, For comparison, chlorophyll fluorescence sections are also presented for the second (**f**) and third (**g**) surveys. In panels **a**, **f** and **g**, density contour lines are overlaid, and inset maps indicate the relevant SeaSoar³⁰ survey leg plotted in red.

Table 1 | Diatom, flagellate and total carbon biomass in the vicinity of the IFF

Station	Site	Position	Date	Diatom carbon (mg m ⁻³)	Flagellate carbon (mg m ⁻³)	Total carbon (mg m ⁻³)
N (14066) leg 1	North of IFF	64° 47.0' N 08° 30.0' W	25 May 2001	4.37	37.98	56.31
J (14071) leg 1	IFF	64° 35.0' N 10° 41.5' W	26 May 2001	89.88	31.21	138.65
14099 pre-survey 1	IFF	64° 35.0' N 10° 40.0' W	4 June 2001	59.38	67.57	152.67
14108 pre-survey 2	South of IFF	63° 41.0' N 10° 05.0' W	10 June 2001	1.49	22.2	47.44
14127 pre-survey 3	IFF	64° 50.0' N 10° 19.0' W	15 June 2001	31.99	40.99	97.39

accounting for just 8% of the autotrophic carbon (~4 mg C m⁻³): conditions typical of post-bloom community succession. Si(OH)₄ concentrations were low (0.3–0.6 mmol m⁻³). Si(OH)₄ uptake kinetics of the remnant (near-surface) diatom population were determined by observing their response to artificially increasing Si(OH)₄ concentrations²² during on-deck incubations. This population was still viable (Fig. 3) because, although the undisturbed uptake rate (ρSi) was low (0.008 mmol m⁻³ h⁻¹), the specific uptake rates of Si(OH)₄ responded rapidly to increased Si(OH)₄ concentration. The spring diatom bloom had not been maintained at this location.

Station J (26 May 2001), representative of the IFF frontal jet (Fig. 1), also showed low concentrations of surface Si(OH)₄ (~0.6 mmol m⁻³) in a post-diatom-bloom state²². Si(OH)₄ uptake rates were 0.010–0.014 mmol m⁻³ h⁻¹ (Fig. 3). However, the diatom population (Table 1) represented 64% of the total biogenic carbon (~90 mg C m⁻³), more typical of a spring diatom bloom. Such low concentrations of Si(OH)₄ are considered to limit diatom growth^{8,13}. We therefore investigated how this frontal diatom bloom might be prolonged, and the vertical distribution of phytoplankton provided a clue.

Three high-resolution repeated SeaSoar surveys²⁰ showed that the IFF frontal bloom was not a short-lived transient state, but persisted at least from 23 May 2001 (Fig. 1) to 17 June 2001 (Fig. 2). Signals of downwardly transported phytoplankton were clearly visible to over 200 m—over four times the euphotic depth²² and extending considerably below the seasonal pycnocline (density contours, Fig. 2). Photosynthetic efficiency (*F*_v/*F*_m), as measured by fast repetition rate fluorometry²¹, showed that these deep phytoplankton were still healthy. Although direct gravitational sinking of phytoplankton cannot be discounted, these signals are indicative of baroclinic

instability of the IFF²³. In the atmosphere, this process drives weather scales. In the ocean, this process creates eddies and meanders in frontal boundaries, the motion of which is driven by the release of potential energy^{23,24}, and is represented here by the exchange of cold dense Arctic waters descending across the front under warm light Atlantic waters. Unstable eddies and fronts provide a dual mechanism for enhancing carbon export through the direct transport of biomass downwards out of the surface mixed-layer^{25,26}, as well as refertilization of an exhausted mixed layer by upwelling^{1,27}.

Currents associated with unstable fronts are difficult to measure. In particular, vertical velocities, typically tens of metres per day, can only be diagnosed by applying a balance condition to the measured horizontal velocities^{23,26,28} (see Methods). Although the vertical velocities seem small, over the course of a day they can transport phytoplankton-rich surface waters downwards out of the euphotic zone, and nutrient-rich deep waters upwards into the euphotic zone.

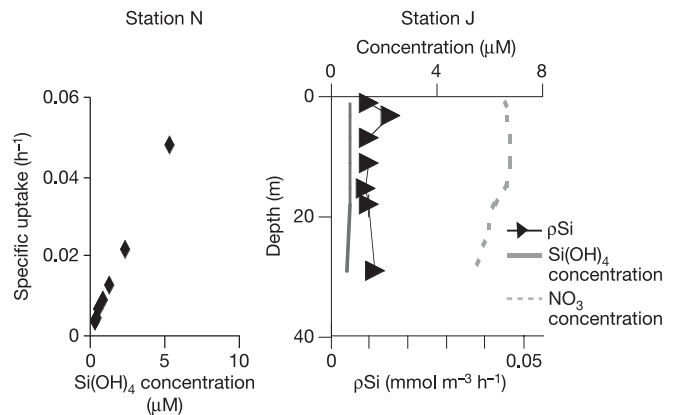


Figure 3 | The results of Si(OH)₄ uptake incubation experiments. Left panel, silicate uptake response to artificially increased Si(OH)₄ concentrations at station N, north of the IFF. Right panel, Si(OH)₄ concentration, nitrate concentration and Si(OH)₄ uptake rates (ρSi) for station J, in the IFF frontal jet.

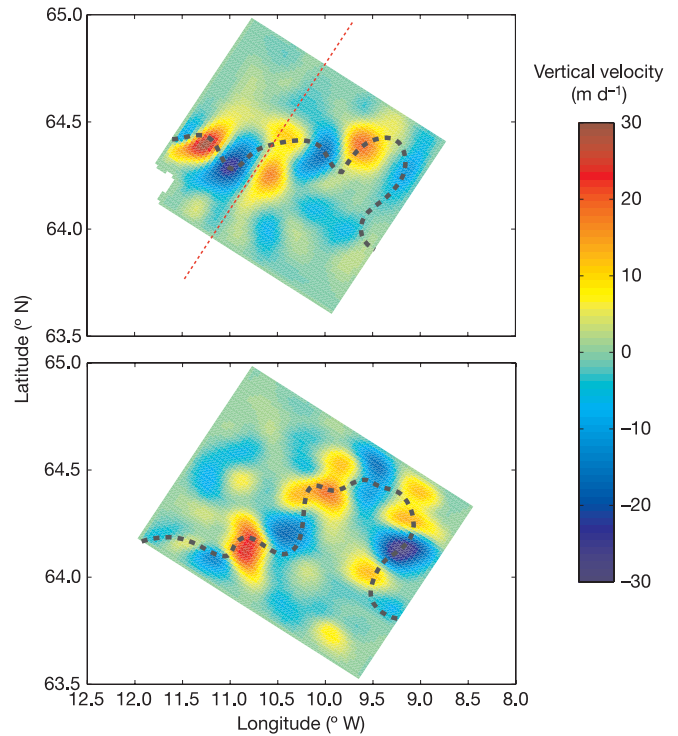


Figure 4 | Diagnosed (QG) vertical motion (positive upwards). Only SeaSoar³⁰ surveys 1 (top panel) and 3 (bottom panel) are shown for the sake of brevity. The vertical velocity is presented in the form of maps on the 1027.75 kg m⁻³ potential density surface, which represents the main pycnocline (Fig. 2). The dotted grey lines mark the position of the front as defined by a depth of ~200 m for the 1027.75 kg m⁻³ density surface. The dotted red line in survey 1 (top panel) marks the vertical velocity section in Fig. 2.

The three SeaSoar surveys provided the 12-day-evolution of the diagnosed vertical velocity field (Fig. 4). The ubiquity of the motion is clearly visible, with peak speeds in excess of 30 m d^{-1} .

Targeted CTD (conductivity–temperature–depth) and phytoplankton taxonomy sampling stations between each SeaSoar survey²⁰ showed that by the end of the FISHES experiment, the IFF still supported high concentrations of diatoms ($\sim 32 \text{ mg C m}^{-3}$ in the IFF compared to $\sim 1.5 \text{ mg C m}^{-3}$ away from the front) against an increasing *Phaeocystis* biomass (Table 1). The $^{32}\text{Si}(\text{OH})_4$ uptake rates measured would exhaust the $\text{Si}(\text{OH})_4$ within a horizontally advected distance of just 60–100 km along the frontal jet. As we emphasized earlier, current knowledge suggests that the observed concentrations of $\text{Si}(\text{OH})_4$ would already be limiting for diatom growth²². Thus, horizontal advection would not result in the ocean colour-signal presented in Fig. 1. Alternatively, if the diatom bloom were solely supported by $\text{Si}(\text{OH})_4$ remineralization in surface layers, then a diatom bloom population should also have been maintained north of the front; however, this was not the case. Thus, by inference, we conclude that vertical transport resulting from baroclinic instability of the IFF is the most probable mechanism for $\text{Si}(\text{OH})_4$ refertilization of the euphotic zone, thereby prolonging the diatom bloom.

We then assessed whether this proposal would allow sufficient resources for diatom production. $\text{Si}(\text{OH})_4$ concentrations²² below the surface layer were $\sim 5 \text{ mmol m}^{-3}$. Upward vertical velocities (Fig. 4) in the order of 10 m d^{-1} cover $>20\%$ of the area of the frontal jet (see Methods). Therefore, following ref. 27, we calculate a $\text{Si}(\text{OH})_4$ supply rate of $10 \text{ mmol m}^{-2} \text{ d}^{-1}$. Over the 29-m depth of the euphotic surface layer (here defined as the 0.1% surface irradiance depth), this supply is equivalent to $0.34 \text{ mmol m}^{-3} \text{ d}^{-1}$ or $\sim 0.014 \text{ mmol m}^{-3} \text{ h}^{-1}$, which agrees well with the measured $\text{Si}(\text{OH})_4$ uptake discussed earlier. Here we assume a 24-h uptake, reflecting our observed decoupling of silica and carbon uptake rates by diatoms, where dark-incubated $\text{Si}(\text{OH})_4$ uptake rates averaged 81% of light-incubated uptake rates²².

At station J, in the IFF, the ratio of particulate organic carbon (POC) to biogenic silica was 9.8, averaged over the euphotic depth. Using standard ^{14}C -incubation techniques²², the carbon-to- $\text{Si}(\text{OH})_4$ uptake ratio was determined to be ~ 9 , assuming both light and dark $\text{Si}(\text{OH})_4$ uptake. These ratios are higher than the 7.5:1 ratio that is conventionally assumed²², and reflect the 64% diatom biomass composition indicated previously. The measured POC:chlorophyll ratio was 36:1, averaged over the euphotic depth. Thus, a $\text{Si}(\text{OH})_4$ supply of $0.010\text{--}0.014 \text{ mmol m}^{-3} \text{ h}^{-1}$ would support new chlorophyll production of $0.8\text{--}1.1 \text{ mg m}^{-3} \text{ d}^{-1}$, and hence the ocean colour-signal in Fig. 1. Over the 25 days of our observations, the IFF resupplied 8.5 mmol m^{-3} of new $\text{Si}(\text{OH})_4$, doubling the total export potential when compared with the annual $\text{Si}(\text{OH})_4$ pool ($6\text{--}8 \text{ mmol m}^{-3}$ discussed earlier).

Thus, diatoms exploit the continued (but not continuous) $\text{Si}(\text{OH})_4$ supply provided by physical processes at an unstable ocean front. Although insufficient to initiate a bloom, this periodic $\text{Si}(\text{OH})_4$ supply prolongs a pre-existing bloom against grazing pressures and species competition. By opportunistically ‘mining’ the $\text{Si}(\text{OH})_4$ supply and exporting carbon from surface layers through mortality, diatoms form a natural ‘chain-gang’ that enhances carbon sequestration in frontal regions of the northeast Atlantic.

Observations of sufficiency, where supply and demand are closely matched, are not expected to give large signals in nutrient concentrations or uptake rates. By directly comparing diagnosed vertical velocities with surface nutrients in Fig. 2, we can speculate on the transient upwelling of a pool of $\text{Si}(\text{OH})_4$ not yet used by diatoms. However, the state observations result from transport over the preceding 24–48 h, and the vertical transport provides a predominantly across-front water exchange in an along-front jet. Direct convolution of the diagnosed vertical motion with state parameters

is therefore not reliable²⁶, hence the requirement for the scaling analysis presented here. Finally, we cannot discount differences in sinking and remineralization rates across the frontal region, and their significance requires further investigation.

Globally, the mechanism presented here and predicted from modelling studies¹ is ubiquitous and effects a vertical exchange of water across ocean fronts. Enhanced export from similarly prolonged new production may be expected in nitrate- (for example, in the Sargasso Sea), phosphate- (for example, in the Mediterranean) or iron- (for example, in the Southern Ocean) limited frontal regions.

METHODS

$\text{Si}(\text{OH})_4$ uptake experiments. At station J, water samples were collected from depths equivalent to 97, 45, 18, 8, 3, 1 and 0.1% of surface PAR (photosynthetically available radiation). The samples were spiked with tracer amounts of ^{32}Si , a high specific-activity radioactive isotope (specific activity $28 \text{ kBq } \mu\text{g}^{-1}$), which allowed the rapid and precise estimation of $\text{Si}(\text{OH})_4$ uptake by phytoplankton^{22,29}. The samples were placed in on-deck incubators, covered with filters to simulate *in situ* underwater irradiance, and cooled with continuously flowing subsurface sea water. Each experiment was carried out from dawn to midday to simulate a half-day cycle, and was terminated by filtering the sample onto a $0.8 \text{ } \mu\text{m}$ -polycarbonate filter. The particulate biogenic Si was digested and treated with a chemical scintillation cocktail, and the ^{32}Si taken up during the incubations was assayed using a liquid scintillation counter^{22,29}. At station N, an uptake kinetics study was carried out²². Here, a single near-surface water sample was collected and divided into eight sub-samples. Incremental quantities of a $1.5 \text{ } \mu\text{M}$ NaSiO_3 solution were added to increase the dissolved silica concentration in seven of the sub-samples. The sub-samples were then spiked with ^{32}Si tracer, incubated and processed as described above²².

Vertical velocity diagnoses. Vertical velocities were diagnosed by applying a balance condition to the measured horizontal velocities^{23,26,28}. This involves solving the quasi-geostrophic omega equation

$$f^2 \frac{\partial^2 w}{\partial z^2} + \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) (N^2 w) = \nabla_h \cdot \mathbf{Q}$$

where x , y and z are the usual right-handed axis set with z positive upwards, f is the local Coriolis parameter on the rotating earth and w is the vertical component of the total velocity field. N is the Brunt-Väisälä frequency, the buoyancy response to a vertical perturbation, and thus determined by the stratification. $\mathbf{Q}$ is a vector that depends only on the gradient of the pressure-driven (geostrophic) horizontal velocity field in all three space dimensions: the geostrophic velocity field is simply that determined by balancing the horizontal pressure gradient with the Coriolis parameter.

Scaling of vertical velocities in the frontal zone. From the ocean colour-satellite image (Fig. 1), the typical width of high-chlorophyll filaments is estimated as $0.1\text{--}0.2$ degrees of latitude ($\sim 10\text{--}20 \text{ km}$). From Fig. 2, the maximum slope of density surfaces (for example, $1027.75 \text{ kg m}^{-3}$) on any cross-front section is $50\text{--}100 \text{ m}$ in $0.1\text{--}0.2$ degrees of latitude. In Fig. 4, the frontal boundary at a depth of 200 m was defined as the position of the $1027.75 \text{ kg m}^{-3}$ density contour. Considering all points within (1) a 50-m depth range around 200 m (that is, $175\text{--}225 \text{ m}$) on this density surface, then 25% of these points have diagnosed vertical velocities in excess of 10 m d^{-1} , and (2) a 100-m depth range around 200 m (that is, $150\text{--}250 \text{ m}$) on this density surface, then 15% of these points have diagnosed vertical velocities in excess of 10 m d^{-1} .

Received 1 February; accepted 20 June 2005.

- Lévy, M., Klein, P. & Tréguer, A.-M. Impact of sub-mesoscale physics on production and subduction of phytoplankton in an oligotrophic regime. *J. Mar. Res.* **59**, 535–565 (2001).
- Savidge, G., Boyd, P., Pomroy, A., Harbour, D. & Joint, I. Phytoplankton production and biomass estimates in the north-east Atlantic Ocean, May–June 1990. *Deep-Sea Res.* **42**, 599–617 (1995).
- Yallop, M. L. Distribution patterns and biomass estimates of diatoms and autotrophic dinoflagellates in the NE Atlantic during June and July 1996. *Deep-Sea Res.* **48**, 825–844 (2001).
- Koeve, W. Wintertime nutrients in the North Atlantic—New approaches and implications for new production estimates. *Mar. Chem.* **74**, 245–260 (2001).
- Ducklow, H. W. & Harris, R. P. Introduction to the JGOFS North-Atlantic Bloom Experiment. *Deep-Sea Res.* **40**, 1–8 (1993).
- Martin-Jezequel, V., Hildebrand, M. & Brzezinski, M. A. Silicon metabolism in diatoms: Implications for growth. *J. Phycol.* **36**, 821–840 (2000).
- Honjo, S. & Mangani, S. J. Annual biogenic particle fluxes to the interior of the North Atlantic Ocean studied at $34^\circ \text{ N } 21^\circ \text{ W}$ and $48^\circ \text{ N } 21^\circ \text{ W}$. *Deep-Sea Res.* **40**, 587–607 (1993).

8. Egge, J. K. & Aksnes, D. L. Silicate as regulating nutrient in phytoplankton competition. *Mar. Ecol. Prog. Ser.* **83**, 281–289 (1992).
9. Sieracki, M. E., Verity, P. G. & Stoecker, D. K. Plankton community response to sequential silicate and nitrate depletion during the 1989 North Atlantic spring bloom. *Deep-Sea Res. II* **40**, 213–225 (1993).
10. Sverdrup, H. U. On the conditions for the vernal blooming of phytoplankton. *J. Cons. Perm. Int. Explor. Mer.* **18**, 287–295 (1953).
11. Longhurst, A. R. *Ecological Geography of the Sea* (Academic, San Diego, 1998).
12. Dugdale, R. C., Wilkerson, F. P. & Minas, H. J. The role of a silicate pump in driving new production. *Deep-Sea Res. I* **42**, 697–719 (1995).
13. Yool, A. & Tyrell, T. Role of diatoms in regulating the ocean's silicon cycle. *Glob. Biogeochem. Cycles* **17**, 1103 (2003) (doi:10.1029/2002GB002018).
14. Lovejoy, C., Legendre, L. & Price, N. M. Prolonged diatom blooms and microbial food web dynamics: experimental results from an Arctic polynya. *Aquat. Microb. Ecol.* **29**, 267–278 (2002).
15. Dugdale, R. C. & Wilkerson, F. P. Silicate regulation of new production in the equatorial Pacific upwelling. *Nature* **391**, 270–273 (1998).
16. Brzezinski, M. A., Villareal, T. A. & Lipschultz, F. Silica production and the contribution of diatoms to new and primary production in the central North Pacific. *Mar. Ecol. Prog. Ser.* **167**, 89–104 (1998).
17. Louanchi, F. & Najjar, R. G. Annual cycles of nutrients and oxygen in the upper layers of the North Atlantic Ocean. *Deep-Sea Res. II* **48**, 2155–2171 (2001).
18. Allen, J. T., Smeed, D. A. & Chadwick, A. L. Eddies and mixing at the Iceland Faeroes Front. *Deep-Sea Res. I* **41**, 51–79 (1994).
19. Strass, V. H. Chlorophyll patchiness caused by mesoscale upwelling at fronts. *Deep-Sea Res.* **39**, 75–96 (1992).
20. Allen, J.T. et al. RRS *Discovery* Cruise 253, 04 May–20 June 2001. Faeroes–Iceland–Scotland Hydrographic and Environmental Survey (FISHES). Southampton Oceanography Centre Cruise Report No. 37, 1–206 (<http://eprints.soton.ac.uk/245/>) (2001).
21. Moore, C. M., Lucas, M. I., Sanders, R. & Davidson, R. Basin-scale variability of phytoplankton bio-optical characteristics in relation to bloom state and community structure in the North East Atlantic. *Deep-Sea Res. I* **52**, 401–419 (2005).
22. Brown, L., Sanders, R., Savidge, G. & Lucas, C. H. The uptake of silica during the spring bloom in the north east Atlantic ocean. *Limnol. Oceanogr.* **48**, 1831–1845 (2003).
23. Allen, J. T. & Smeed, D. A. Potential vorticity and vertical velocity at the Iceland Faeroes Front. *J. Phys. Oceanogr.* **26**, 2611–2634 (1996).
24. Munk, W., Armi, L., Fischer, K. & Zachariasen, F. Spirals on the sea. *Proc. R. Soc. Lond. A* **456**, 1217–1280 (2000).
25. Fielding, S. et al. Mesoscale subduction at the Almeria–Oran front. Part 2: Biophysical interactions. *J. Mar. Syst.* **30**, 287–304 (2001).
26. Allen, J. T., Ruiz, S., Smeed, D. A. & Tintoré, J. Mesoscale subduction at the Almeria–Oran front. Part 1: Ageostrophic flow. *J. Mar. Syst.* **30**, 263–285 (2001).
27. Naveira-Garabato, A. C., Strass, V. H. & Kattner, G. Fluxes of nutrients in a three-dimensional meander structure of the Antarctic Polar Front. *Deep-Sea Res. II* **49**, 3771–3792 (2002).
28. Shearman, R. K., Barth, J. A. & Kosro, P. M. Diagnosis of the three-dimensional circulation associated with mesoscale motion in the California Current. *J. Phys. Oceanogr.* **29**, 651–670 (1999).
29. Brzezinski, M. A. & Phillips, D. A. Evaluation of ^{32}Si as a tracer for measuring silica production rates in marine waters. *Limnol. Oceanogr.* **42**, 856–865 (1997).
30. Allen, J., Dunning, J., Cornell, V., Moore, M. & Crisp, N. Operational oceanography using the 'new' SeaSoar ocean undulator. *Sea Technol.* **43**, 35–40 (2002).

Acknowledgements FISHES was a multidisciplinary experiment conducted as part of the Biophysical Interactions and Controls on Export Production core strategic programme. We thank R. Pollard for initiating and developing the FISHES experiment, which fully met its objectives owing to the enthusiasm and professionalism of all the officers, crew and scientists of RRS *Discovery* cruise 253. We thank the Defence Evaluation and Research Agency (now the Defence Science and Technology Laboratory) for their support of a number of staff involved in FISHES through the development of the Techniques for Optimising Environmental Sampling Joint Grant Scheme project. We thank the Natural Environment Research Council for their support of the George Deacon Division, and the National Oceanography Centre for their commitment to core strategic science funding.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.T.A. (jta@noc.soton.ac.uk).

Meniscus-climbing insects

David L. Hu¹ & John W. M. Bush¹

Water-walking insects and spiders rely on surface tension for static weight support^{1,2} and use a variety of means to propel themselves along the surface^{3–8}. To pass from the water surface to land, they must contend with the slippery slopes of the menisci that border the water's edge. The ability to climb menisci is a skill exploited by water-walking insects as they seek land in order to lay eggs or avoid predators⁴; moreover, it was a necessary adaptation for their ancestors as they evolved from terrestrials to live exclusively on the water surface³. Many millimetre-scale water-walking insects are unable to climb menisci using their traditional means of propulsion^{2,3,9}. Through a combined experimental and theoretical study, here we investigate the meniscus-climbing technique that such insects use. By assuming a fixed body posture, they deform the water surface in order to generate capillary forces^{10–13}; they thus propel themselves laterally without moving their appendages. We develop a theoretical model for this novel mode of propulsion and use it to rationalize the climbers' characteristic body postures and predict climbing trajectories consistent with those reported here and elsewhere³.

Although the surfaces of standing bodies of water such as ponds are flat on human scales, there is significant topography on the scale of millimetre-scale water-walking insects. To water-walking insects small relative to the capillary length ($l_c = (\sigma/\rho g)^{1/2} = 0.27$ cm, where $\sigma = 70$ dynes cm⁻¹ is the surface tension, $\rho = 1$ g cm⁻³ the water density, and $g = 980$ cm s⁻² the gravitational acceleration), the menisci adjoining land and emergent vegetation appear as frictionless mountains. Some water-walking arthropods, such as the adult water strider³ and fisher spider¹⁴, are capable of leaping over menisci. Some insects cannot do this, but have developed an ingenious means of ascent. By locking into a fixed posture in which they deform the free surface, they generate a tangential force that draws them up the slope (Fig. 1). Meniscus climbing has been reported and described qualitatively by previous investigators^{2,3,9}; however, the precise physical mechanism was not elucidated and no theoretical model was proposed.

Fluid mechanicians^{10–12,15} have long known that lateral capillary forces exist between small floating objects, an effect responsible for the formation of bubble rafts in champagne and the clumping of breakfast cereal in a bowl of milk¹⁶. Recently, such capillary forces have been recognized as a means by which to promote self-assembly of microstructures at a free surface^{13,17–19}. Here we describe the use of analogous capillary effects in the ascent of menisci by semiaquatic insects.

We collected water-walking insects *Mesovelia* and *Microvelia* and beetle larvae *Pyrrhalta* from local freshwater ponds and maintained them in aquaria. The insects were filmed using a high-speed video camera (500 frames s⁻¹) and the images digitized and analysed using Midas motion analysis software.

An insect approaching an object long relative to the capillary length, such as a shoreline or floating log, must first ascend an effectively two-dimensional meniscus whose shape, $z = \eta(x)$, is prescribed by the Young–Laplace equation²⁰, $\rho g \eta = \sigma \nabla \cdot \mathbf{n}$ (where $\mathbf{n}$

is the unit outward normal; see Fig. 1c), that expresses the balance between hydrostatic and curvature pressures at the interface. Provided the interfacial slope is everywhere sufficiently small, $\eta_x \ll 1$, this equation may be solved to yield a meniscus shape of $\eta(x) = l_c \cot \theta e^{-x/l_c}$ where θ is the contact angle¹⁰. Consider a small massless body that deforms the surface by applying a vertical force $T\hat{z}$ at a horizontal distance x_0 from the wall. Provided the body is small relative to the capillary length, the applied force may be treated as a point force and the lateral force on the body may be written exactly¹¹ as

$$\mathbf{F}(x_0) = -T \cot \theta e^{-x_0/l_c} \hat{x} \quad (1)$$

A small buoyant object ($T > 0$) floating in the presence of a larger meniscus is thus drawn up the slope¹²; conversely, a negatively buoyant body ($T < 0$) is driven downwards.

Water-walking insects are generally covered with a mat of dense hair that renders them effectively non-wetting^{4,21}, but some have developed specialized feet with retractable hydrophilic claws or 'ungues' that allow them to raise the free surface^{2,4}, an adaptation critical for meniscus-climbing. Water-walking insects such as *Mesovelia* (Fig. 1) ascend menisci by assuming a static posture in which they pull up on the free surface with the wetting tips of their front and rear tarsi and push down with their middle tarsi.

A simple theoretical model allows us to rationalize this body configuration. Consider an insect of mass M with its centre of mass a horizontal distance x_0 from a vertical wall (Fig. 1c). Each of its front and rear tarsi pull up with forces F_1 and F_3 at horizontal positions $x_0 - L_1$ and $x_0 + L_3$, respectively; the middle tarsi push down with force F_2 at the position $x_0 + L_2$. According to equation (1), the front and rear legs are attracted to the background meniscus while the middle pair are repelled. The insect may adjust L_i subject to the geometrical constraints imposed by its size, and F_i subject to the constraint that the force per unit length applied on any of its limbs cannot exceed 2σ , lest its legs pass through the free surface.

The local slope of the interface is assumed to be small and the forces F_i applied normal to the meniscus. The normal force balance on the insect and torque balance about the middle legs are written in the Methods section as equations (2) and (3), respectively. The acceleration of the insect is determined by the tangential force balance in equation (4). We note that if the insect is too heavy, it will be unable to climb owing to the combined influence of its weight and the repulsive capillary force generated by its middle legs. Because the capillary force decays exponentially with distance from the wall, the insect generates the majority of its thrust with its front legs. It should thus extend its front legs and pull up as hard as possible, while maximizing the extension of its back legs and pulling up with the force sufficient to balance torques, as prescribed by equation (3). Because the range of capillary forces corresponds to the capillary length, insects must scramble onto the meniscus to initiate their ascent.

We have thus far assumed for the sake of simplicity that the meniscus climber assumes a laterally symmetric leg configuration.

¹Department of Mathematics, MIT, Cambridge, Massachusetts 02139, USA.

Andersen³ and Miyamoto⁹ note that several climbers, such as *Hydrometra*, tilt their bodies during ascent (see inset B in Fig. 3). Presumably, such tilted postures are assumed in order to maximize the capillary thrust. Theoretical modelling of the tilted postures requires consideration of additional torque balances that are immediately satisfied in the symmetric posture, specifically, those that prevent the insect's angular acceleration about a vertical axis (yaw) and about an axis aligned with the mean direction of motion (roll).

Miyamoto⁹ reported that a number of terrestrial insects have also developed the ability to ascend menisci, an adaptation exploited as they seek land having fallen onto water, often from overhanging vegetation. Unlike water-walking insects whose hairy legs render them effectively non-wetting^{3,20,21}, terrestrial insects must deform the surface with their wetting body perimeters. For example, the larva of the waterlily leaf beetle is circumscribed by a contact line, and deforms the free surface by arching its back (Fig. 2). The beetle larva will be drawn up the meniscus if the anomalous surface energy

generated by arching its back exceeds the gain in the system's gravitational potential energy associated with its ascent. By arching its back to match the curvature of the meniscus, the beetle larva may generate a lateral force that drives it up the meniscus. We note that for a randomly oriented beetle larva, the forces (in equation (1)) produced by its arched posture generate a torque about the vertical axis that serves to align the beetle larva perpendicular to the meniscus; the beetle thus abuts the wall tail first.

Figure 3 shows the observed trajectories of a *Mesovelia* individual of length 2 mm and weight 0.2 dynes. Accompanying theoretical trajectories were obtained by numerically integrating equation (4). The insect's position and leg configuration were recorded by high-speed video; the insect's weight and the meniscus contact angle on plexiglass (40°) were measured using a scale and a still camera, respectively. Given the leg position and meniscus contact angle, there is a single unknown in the model, F_1 . For *Mesovelia*, F_1 was inferred from the trajectory to be 2–4 dynes, values comparable to $\sigma\pi w$, the maximum force the front leg tip of diameter $w \approx 60 \mu\text{m}$ can apply to

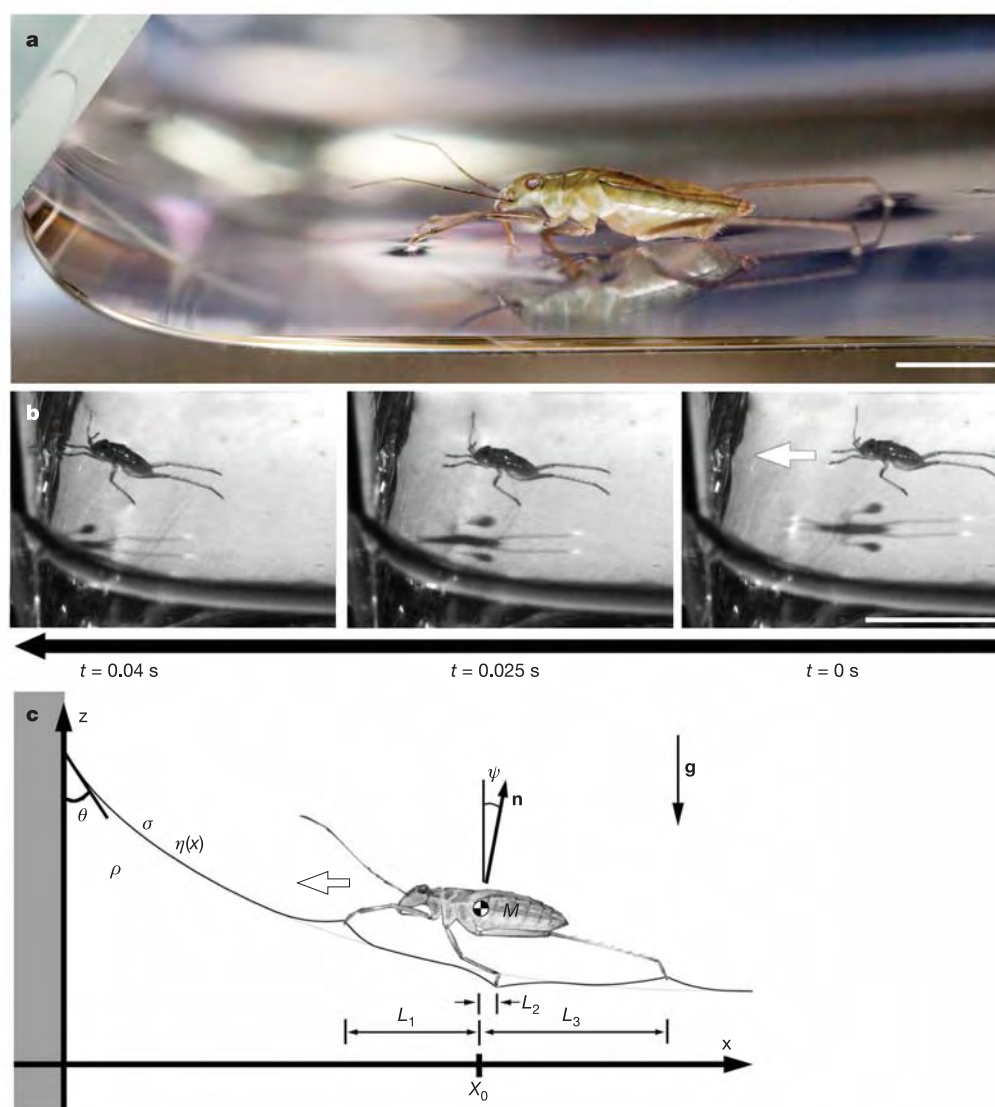


Figure 1 | Meniscus climbing by the water treader *Mesovelia*. **a**, *Mesovelia* approaches a meniscus, from right to left. The deformation of the free surface is evident near its front and hind tarsi. **b**, High-speed video images of an ascent. Lighting from above reveals the surface deformation produced. In pulling up, the insect generates a meniscus that focuses the light into a bright spot; in pushing down, it generates a meniscus through which light is

diffused, casting a dark spot. Characteristic speeds are $1\text{--}10 \text{ cm s}^{-1}$. Scale bars, 3 mm. See Supplementary Information for the accompanying video sequence. **c**, Schematic illustration of the meniscus-climbing *Mesovelia*: it pulls up with its wetting front and hind claws, and pushes down with its middle legs. $\mathbf{n}$ denotes the normal to the undeformed meniscus, and x_0 the lateral position of the insect's centre of mass.

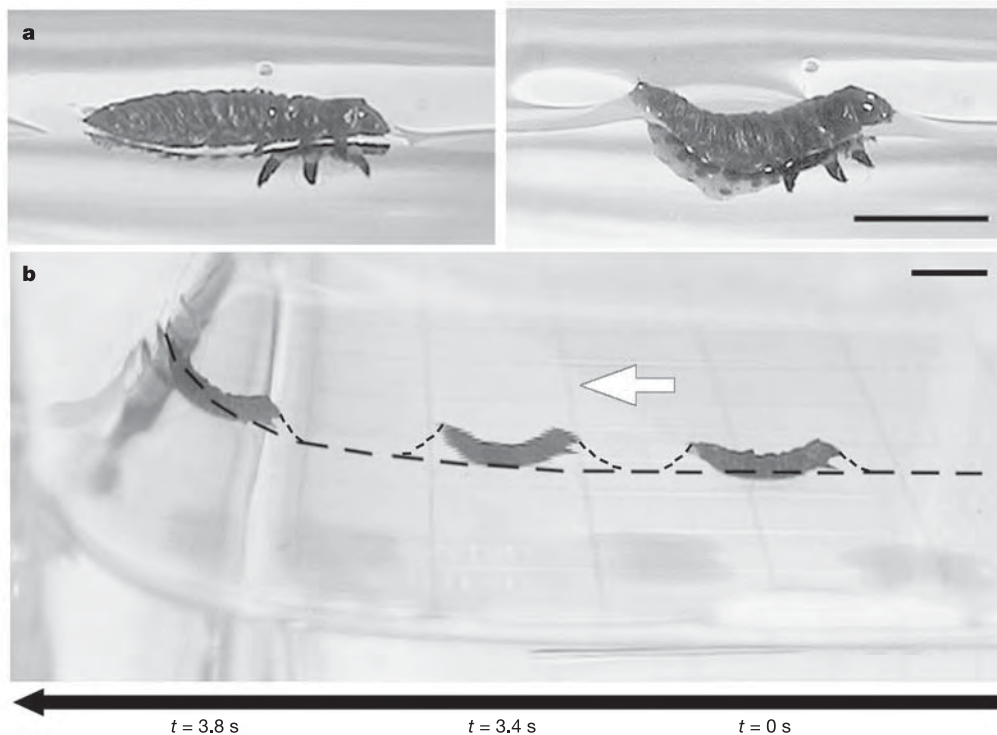


Figure 2 | Meniscus climbing by the larva of the waterlily leaf beetle. The beetle larva is a terrestrial insect unsuited to walking on water but it is nevertheless able to ascend the meniscus, from right to left. **a**, It is partially wetting and so circumscribed by a contact line. It deforms the free surface by

arching its back, thus generating the desired capillary thrust. **b**, The beetle larva's ascent is marked by peak speeds in excess of 10 cm s^{-1} . Scale bars, 3 mm. See Supplementary Information for the accompanying video sequence.

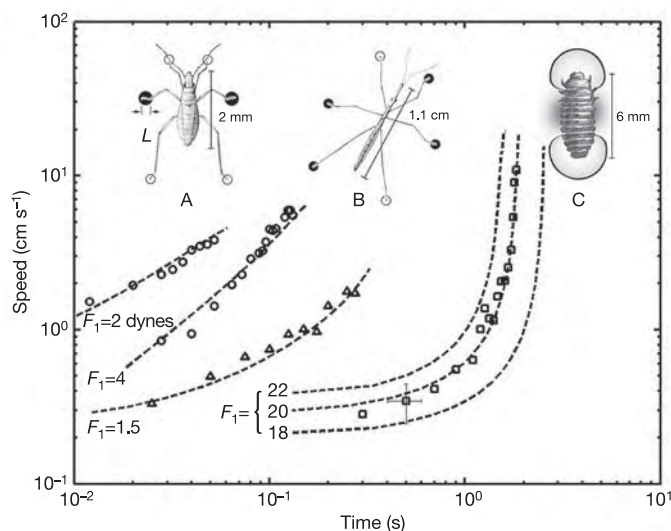


Figure 3 | Observed evolution of insect speed during the ascent of menisci. Insects are *Mesovelia* (A, circles), *Hydrometra*³ (B, triangles) and the beetle larva (C, squares). Circles around the insects' feet indicate the sense of the surface deflection: white upwards and black downwards. Theoretical predictions based on equation (4) are presented as dashed lines. Inferred forces F_1 applied by the frontmost appendage are given in dynes. The sensitivity of the theoretical trajectories to F_1 are indicated for the beetle larva trajectory. For the *Hydrometra* data³, the best-fit trajectory yielded the net capillary force, from which the individual F_i values were calculated. We assumed the meniscus contact angle to be 40° , comparable to the contact angle of water on plexiglass, and then inferred the net capillary force by optimally fitting the observed trajectory. The individual forces F_i applied by the insect were then calculated by consideration of the four governing equations (the normal force balance and three torque balances) and the upper bounds on the forces applied by each leg. The error bars (shown for one point but applicable to every data point) reflect the uncertainty associated with measuring the body position.

the free surface without breaking through. The corresponding force F_2 is comparable to the maximum force, $2\sigma L$, the insect leg can apply without breaking through with the tarsal segment of length L of its middle leg (see inset A in Fig. 3). The trajectory of the beetle larva, of length 6 mm and weight 150 dynes, was similarly computed using equation (4), by setting $F_2 = 0$. Again, the single unknown in modelling the beetle larva's ascent is F_1 , which is equal to F_3 by symmetry; the best fit is obtained with a force of 20 dynes that corresponds to the maximum surface tension force it can generate with its frontal perimeter of 3 mm. Figure 3 also shows the climbing trajectory for the tilting climber *Hydrometra*, a water-walker of length 1.1 cm and weight 1.8 dynes, reported by Andersen³. An accompanying theoretical trajectory was computed from the reported leg positions and inferred values of θ and the net capillary force.

Meniscus climbing is an unusual means of propulsion in that the insect propels itself in a quasi-static configuration, without moving its appendages. Biocomotion is generally characterized by the transfer of muscular strain energy to the kinetic and gravitational potential energy of the creature, and the kinetic energy of the suspending fluid^{22–25}. In contrast, meniscus climbing has a different energy pathway: by deforming the free surface, the insect converts muscular strain to the surface energy that powers its ascent.

METHODS

The following equations were considered in computing the trajectories of the meniscus-climbing insects. Normal and tangential directions refer to orientation with respect to the meniscus at the point x_0 . We assume that the meniscus slope does not vary appreciably over the length of the insect. Terms are defined in Fig. 1c. The normal force balance on the insect is

$$2F_2 = 2F_1 + 2F_3 + Mg \cos \psi(x_0) \quad (2)$$

The torque balance about the insect's middle legs is

$$2F_3 L_3 = 2F_1 L_1 - Mg L_2 \cos \psi(x_0) \quad (3)$$

The tangential force balance on the insect is

$$M \frac{d^2 x_0}{dt^2} = -2B(F_i, L_i) \cot \theta \cos \psi(x_0) e^{-x_0/l_c} + Mg \sin \psi(x_0) \quad (4)$$

where the first term on the right-hand side represents the net capillary force and

$$B(F_i, L_i) = F_1 e^{L_1/l_c} - F_2 e^{-L_2/l_c} + F_3 e^{-L_3/l_c} \quad (5)$$

As a caveat, we note that the validity of equation (1) and therefore of equation (4) is expected to break down when the slope of the meniscus becomes large.

Received 25 May; accepted 4 July 2005.

- Brocher, F. Les phénomènes capillaires. Leur importance dans la biologie aquatique. *Ann. Biol. Lacustre* **4**, 89–139 (1910).
- Baudoin, R. La physico-chimie des surfaces dans la vie des Arthropodes aériens des miroirs d'eau, des rivages marins et lacustres et de la zone intercotidale. *Bull. Biol. Fr. Belg.* **89**, 16–164 (1955).
- Andersen, N. M. A comparative study of locomotion on the water surface in semiaquatic bugs (Insecta, Hemiptera, Gerromorpha). *Vidensk. Meddr. Dansk. Naturh. Foren.* **139**, 337–396 (1976).
- Andersen, N. M. *The Semiaquatic Bugs (Hemiptera, Gerromorpha): Phylogeny, Adaptations, Biogeography and Classification* (Scandinavian Science, Klampenborg, Denmark, 1982).
- Suter, R. B., Rosenberg, R. B., Loeb, S., Wildman, H. & Long, J. H. Locomotion on the water surface: Propulsive mechanisms of the fisher spider *Dolomedes triton*. *J. Exp. Biol.* **200**, 2523–2538 (1997).
- Suter, R. B. & Wildman, H. Locomotion on the water surface: hydrodynamic constraints on rowing velocity require a gait change. *J. Exp. Biol.* **202**, 2771–2785 (1999).
- Hu, D. L., Chan, B. & Bush, J. W. M. The hydrodynamics of water strider locomotion. *Nature* **424**, 663–666 (2003).
- Bush, J. W. M. & Hu, D. L. Walking on water: Biocomotion at the interface. *Annu. Rev. Fluid Mech.* **38**, 339–369 (in the press).
- Miyamoto, S. On a special mode of locomotion utilizing surface tension at the water-edge in some semiaquatic insects. *Kontyû* **23**, 45–52 (1955).
- Kralchevsky, P. A. & Denkov, N. D. Capillary forces and structuring in layers of colloid particles. *Curr. Opin. Coll. Interf. Sci.* **6**, 383–401 (2001).
- Chan, D. Y. C., Henry, J. D. J. & White, L. R. The interaction of colloidal particles collected at fluid interfaces. *J. Coll. Interf. Sci.* **79**, 410–418 (1981).
- Nicolson, M. The interaction between floating particles. *Proc. Cambr. Phil. Soc.* **45**, 288–295 (1949).
- Whitesides, G. M. & Grzybowski, B. Self-assembly at all scales. *Science* **295**, 2418–2421 (2002).
- Suter, R. B. & Gruenewald, J. Predator avoidance on the water surface? Kinematics and efficacy of vertical jumping by *Dolomedes* (Araneae Pisauridae). *J. Arach.* **28**, 201–210 (2000).
- Gifford, W. A. & Scriven, L. E. On the attraction of floating particles. *Chem. Eng. Sci.* **26**, 287–297 (1971).
- Vella, D. & Mahadevan, L. The 'Cheerios effect'. *Am. J. Phys.* **73**, 817–825 (2005).
- Grzybowski, B. A., Bowden, N., Arias, F., Yang, H. & Whitesides, G. M. Modeling of menisci and capillary forces from the millimeter to the micrometer size. *J. Phys. Chem. B* **105**, 404–412 (2001).
- Manoharan, V. N., Elsesser, M. T. & Pine, D. J. Dense packing and symmetry in small clusters of microspheres. *Science* **301**, 483–487 (2003).
- Lauga, E. & Brenner, M. P. Evaporation-driven assembly of colloidal particles. *Phys. Rev. Lett.* **93**, 238301 (2004).
- de Gennes, P.-G., Brochard-Wyart, F. & Quéré, D. *Capillarity and Wetting Phenomena: Drops, Bubbles, Pearls and Waves* (Springer, Berlin, 2003).
- Gao, X. & Jiang, L. Water-repellent legs of water striders. *Nature* **432**, 36 (2004).
- Dickinson, M. H. et al. How animals move: An integrative view. *Science* **288**, 100–106 (2000).
- Alexander, R. M. *Principles of Animal Locomotion* (Princeton Univ. Press, Princeton, New Jersey, 2003).
- Biewener, A. A. *Animal Locomotion* (Cambridge Univ. Press, Cambridge, 2003).
- Vogel, S. *Biomechanics: Life's Physical World* (Princeton Univ. Press, Princeton, New Jersey, 2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Kreider for his early contributions, B. Chan for his assistance with the illustrations, L. Mendel for photographing Fig. 1a and MIT's Edgerton Center for access to their high-speed video equipment. We gratefully acknowledge the financial support of the NSF.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.W.M.B. (bush@math.mit.edu).

Conformity to cultural norms of tool use in chimpanzees

Andrew Whiten¹, Victoria Horner^{1,2} & Frans B. M. de Waal²

Rich circumstantial evidence suggests that the extensive behavioural diversity recorded in wild great apes reflects a complexity of cultural variation unmatched by species other than our own^{1–12}. However, the capacity for cultural transmission assumed by this interpretation has remained difficult to test rigorously in the field, where the scope for controlled experimentation is limited^{13–16}. Here we show that experimentally introduced technologies will spread within different ape communities. Unobserved by group mates, we first trained a high-ranking female from each of two groups of captive chimpanzees to adopt one of two different tool-use techniques for obtaining food from the same ‘Pan-pipe’ apparatus, then re-introduced each female to her respective group. All but two of 32 chimpanzees mastered the new technique under the influence of their local expert, whereas none did so in a third population lacking an expert. Most chimpanzees adopted the method seeded in their group, and these traditions continued to diverge over time. A subset of chimpanzees that discovered the alternative method nevertheless went on to match the predominant approach of their companions, showing a conformity bias that is regarded as a hallmark of human culture¹¹.

Owing to logistical and ethical constraints on translocation and other field experiments, social learning in apes has been studied experimentally predominantly in captive populations. Over 30 such experiments in the past 15 years have provided evidence of imitation and other forms of social learning¹⁷. However, the extent to which the cultural interpretation of data from the wild is supported by this work remains contentious^{14–15,18}, principally because all of the

controlled experiments have been restricted to one-to-one learning, typically relying on a human model. The extent to which the social learning documented is sufficient to sustain traditions has thus remained unclear.

Our experiment bridges the gap between population-level studies of wild apes and one-to-one social learning experiments by (1) extending the experimental approach to the group level, (2) focusing on ape-to-ape transmission, and (3) using a powerful ‘two-action’ methodology¹⁷. In this approach, individuals see a given task completed using one of two possible techniques, allowing the extent to which their own subsequent behaviour matches the demonstration to be systematically measured. We studied three groups of chimpanzees: a control group exposed to a new task with no expert present, and two experimental groups, each supplied with a familiar, conspecific expert trained to solve this new task in a different way. Unlike previous attempts to study traditions using a single experimental group^{19–22}, our three-group design allows us to measure the extent to which two quite different techniques are copied sufficiently well to become traditions, with the control condition identifying baseline levels of individual discovery.

We used a naturalistic foraging task that we called the ‘Pan-pipes’ (Fig. 1). Unseen by her group mates, a high-ranking adult female from each of two groups was first trained to recover food from a Pan-pipe apparatus exclusively using one of two new tool-use techniques: ‘Lift’ or ‘Poke’ (Fig. 1a, b). The rest of the group (each $n = 16$) was then allowed to watch their local expert (while their access to the Pan-pipes was prevented) over a seven-day observation

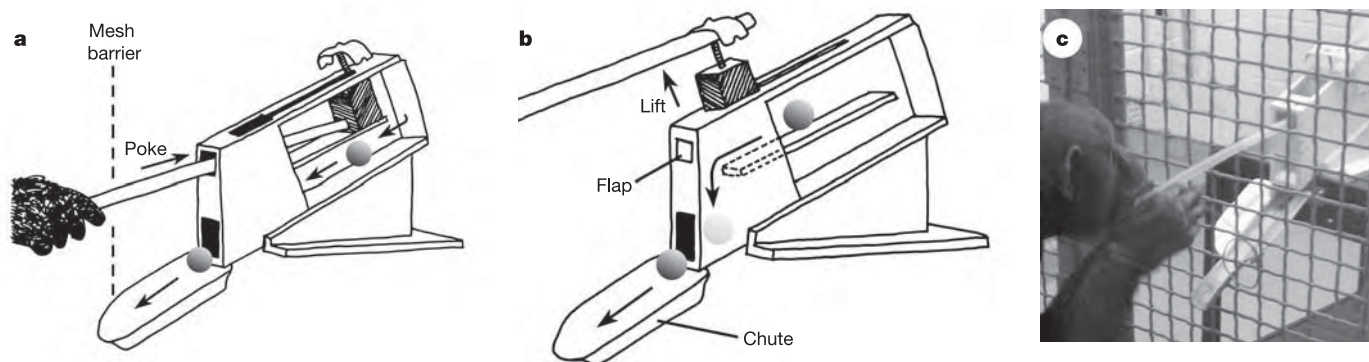


Figure 1 | Two techniques for gaining food from the ‘Pan-pipes’ apparatus. For each technique, the chimpanzee must insert a stick-tool through mesh caging to contact the apparatus and free a desirable food item that is trapped behind a blockage in the upper of two pipes. The food then rolls down a chute into the chimpanzees’ enclosure. The Pan-pipes were 12 cm outside of the enclosure. **a**, In the Poke method, the stick-tool is

inserted under the front flap, pushing the blockage back along the ramp so that the food is knocked off and rolls forward underneath. **b**, In the Lift method, the stick-tool is passed under hooks, allowing the blockage to be lifted and the food to roll forward. **c**, Chimpanzee GG performing the Poke method. A video of Pan-pipe operation can be found in the Supplementary Information.

¹Centre for Social Learning and Cognitive Evolution, School of Psychology, University of St Andrews, St Andrews, Fife, KY16 9JP, UK. ²Living Links, Yerkes National Primate Research Center, Emory University, Atlanta, Georgia 30322, USA.

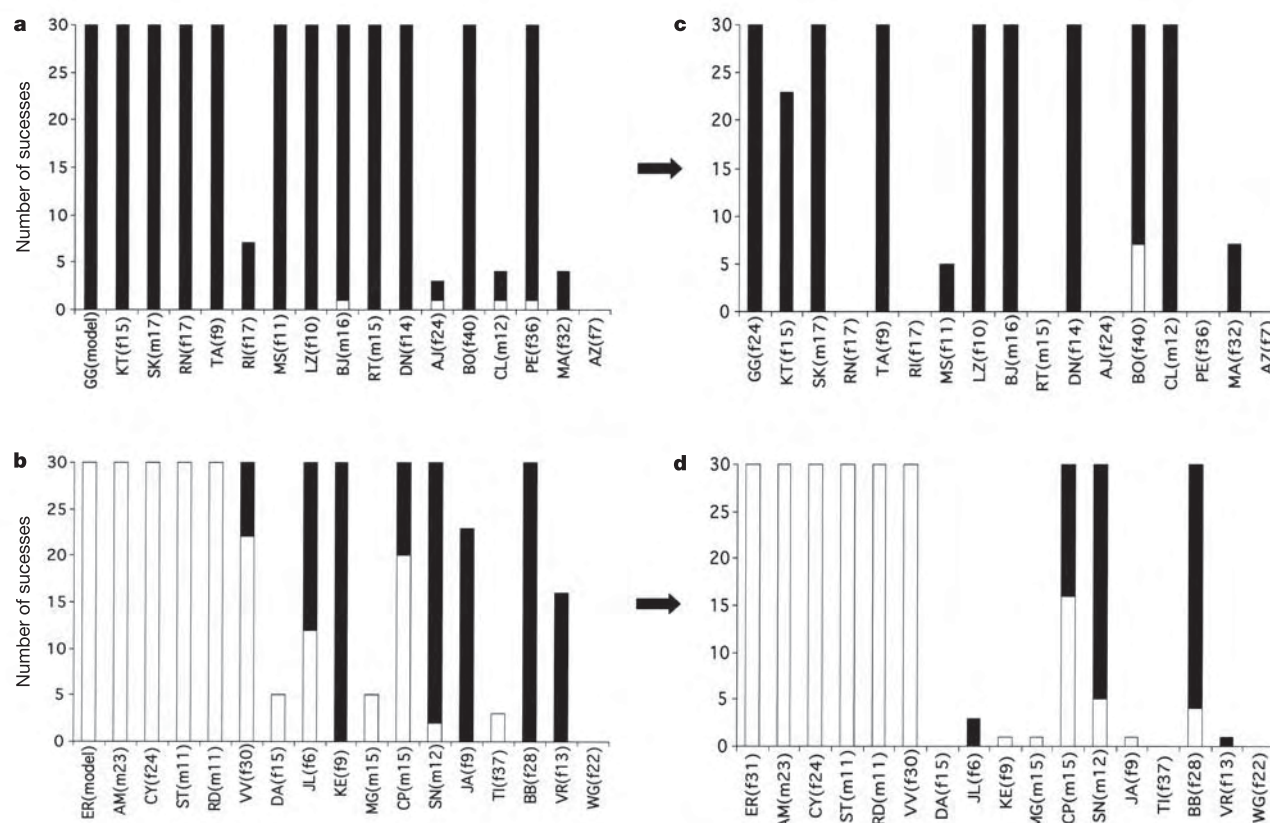


Figure 2 | Adoption of the Poke and Lift methods in two groups. **a**, Poke group at T1. **b**, Lift group at T1. **c**, Poke group at re-test (T2). **d**, Lift group at re-test (T2). Black bars indicate Poke; white bars indicate Lift. The ID code for each chimpanzee is given together with a gender code (m, male; f, female)

and their age in years. Chimpanzees are arranged left to right in the order of their first success in gaining food, starting with the trained model for each group. Data show up to the first 30 successes of each chimpanzee in T1 and T2.

phase of 20-min daily sessions. In this phase, we counted how many demonstrations each chimpanzee paid attention to, which ranged from 8 to 205 per chimpanzee. The expert consistently used the method it had been trained on, and each success was watched by 1–10 chimpanzees (median 4). Both groups showed intense and extended interest in the expert's activities.

The task was then made accessible to all chimpanzees over a total period of 36 h, spread over 10 days (T1). In this phase, 15 chimpanzees in each group successfully used the tool to gain food. Median latency from gaining access to the task until the first success was 29 s (range 5–434 s) for the Poke group, and 52 s (range 2–477 s) for the Lift group. This contrasts with the behaviour of six control chimpanzees offered the task, first individually for 1 h, and then as a group for 4.5 h. All control chimpanzees manipulated the tool and food chute repeatedly, and two inserted the tool into the lower of the Panpipes, but none succeeded in gaining any food (Fisher Exact Test for success, experimental versus control animals, $P < 0.001$).

Preferential adoption of the Lift technique ($\%Lift = \%Lift / (\%Lift + \%Poke)$) in up to the first 30 successes) was significantly greater in the Lift group (median 73.3%, interquartile range 3.3–100.0%) than in the Poke group (median 0.0%, interquartile range 0.0–1.7%) (2-tailed Mann-Whitney U -test, $n_1 = 15$, $n_2 = 15$, $Z = 3.31$, $P < 0.001$) (Fig. 2a, b). In the Poke group, all tool users adopted predominantly the Poke technique. In the Lift group, the first six chimpanzees to succeed adopted the Lift method predominantly. However, chimpanzee JL then discovered both the Poke and Lift techniques, and continued to use both of them (Fig. 2b). Two other chimpanzees in this group then acquired both methods, while two adopted only the Lift method and four only the Poke method.

These data suggest waning of the pattern introduced by the expert, at least in one group. However, when the apparatus was reintroduced

two months later (T2) and data obtained for 23 chimpanzees that worked at the task, the statistically significant difference in the preferential adoption of the Lift method was maintained ($n_1 = 10$, $n_2 = 13$, $Z = 3.56$, $P < 0.001$). Indeed, all but one tool user in the Poke group now used Poke exclusively (Fig. 2c), and all but two in the Lift group used the Lift method, a majority of them exclusively (Fig. 2d). As expected with this naturalistic approach, overall activity at the task varied greatly, constrained by social dynamics and probably by individual characteristics such as motivation and learning capacity.

To our knowledge, these data provide the first robust experimental demonstration of the spread and maintenance of (1) alternative traditions in any primate, and (2) alternative tool-use techniques in any non-human animal. Additionally, the 'two alternatives' methodology shows that learning involves not merely the facilitation of an existing competence, but a capacity to acquire particular local variants of the technique, precisely as required if the behavioural variants identified in wild populations are indeed socially transmitted.

Additionally, we found evidence of a conformist bias, identified in numerous human studies as a powerful tendency to discount personal experience in favour of adopting perceived community norms^{23–25}. As a substantial number of chimpanzees ($n = 14$) succeeded in gaining food using both methods at some point within T1, we were able to define conformity in this context as adoption of the group's norm despite being able to use both methods. Figure 2 suggests that the Poke method was the more conducive to our chimpanzees, so that with no conformity bias, those chimpanzees discovering both methods would be expected to converge on using Poke. Instead, for 13 chimpanzees that used both methods at T1 and also performed at T2, the median percentage of acts matching their

Table 1 | Grades of social conformity

Grade	Characteristics	Chimpanzees in present study
1. Short-latency conformity	After the discovery of an alternative method of task solution, there is nevertheless a rapid convergence on the norm of other group mates (in the present study, this effect was typically observed on the same day).	MS***, BJ***, BO***, PE***, CY***, RD***, VV***
2. Long-latency conformity ('incubation')	After discovery of an alternative method, convergence towards the norm of group mates develops only after an extended period (chimpanzees listed here are those that show an increased proportion of the method consistent with their group norm at the two-month re-test compared with T1).	CL, ST***, SN, CP***, KE, JA, BB**
3. Matching	The norm of group mates is adopted, with no performance of any alternatives. This category could be considered 'blind conformity' insofar as the method conforms to that witnessed, with apparent blindness to alternatives. However, it differs from conformity as defined in the categories above, in which matching occurs despite knowledge of alternative approaches. Thus, this category equates simply to high-fidelity social learning.	KT***, SK***, RN***, TA***, RI**, LZ***, RT***, DN***, MA**, AM***, DA*, MG***, TI*
4. Non-conformity ('contrariness')	A method for task solution contrary to the norm among group mates is exclusively (VR) or increasingly (JL) preferred.	VR, JL

Two forms of strong conformity (short- and long-latency), in which a chimpanzee adopts the group norm despite having discovered an alternative, successful method, are distinguished here from a weaker sense of conformity (matching) that corresponds to high-fidelity social learning. Chimpanzees from the Poke group are shown in bold. Uncorrected binomial and chi-squared statistics are used to indicate the strength of effect for each individual (see main text for group-level statistical tests). Chimpanzees listed in Grade 1 are those that showed a rapid return to their group norm after demonstrating both methods. Three asterisks, $P < 0.001$ (binomial test). All conformed during T1 except for chimpanzee BO (BO performed the alternative method, briefly, only at T2). Chimpanzees listed in Grade 2 showed a shift in bias towards their group norm only from T1 to T2. Two asterisks, $P < 0.05$; three asterisks, $P < 0.001$ (chi-squared test). Chimpanzees in Grade 3 showed an immediate bias to their group norm. Asterisk, $P < 0.05$ (one-tailed); two asterisks, $P < 0.05$ (binomial test); three asterisks, $P < 0.001$ (binomial test). AJ provided too little data to be classified. See Fig. 2 for age and sex of chimpanzees. Raw frequency data are provided in the Supplementary Information.

group norm was significantly higher at T2 than T1 (T1 median 75.0%, interquartile range 11.8–99.4%; T2 median 99.4%, interquartile range 28.7–100%; two-tailed Wilcoxon test, n -ties = 14, $T^+ = 14$, $P < 0.02$), and for eight of these chimpanzees, the increase was in Lift. A strong conformist bias had already been seen in six chimpanzees at T1, when each achieved over 36 successes but the percentage of contrary acts never exceeded 3% in either group (bias to group norm always $P < 0.001$, binomial test) (see Table 1).

These findings encourage us to outline a tentative taxonomy of conformist and allied effects (Table 1). Using this classification, we find that once we have accounted for those chimpanzees in our study that either (1) discover the alternative technique but subsequently show evidence of conformity, or (2) 'conform' in the more basic sense of adopting the local technique without discovering the alternative, only two 'non-conformist' chimpanzees remain, exclusively or increasingly performing the alternative technique. As many as 18 out of 30 of the successful chimpanzees we studied showed substantial conformity effects at the individual level (indicated by triple asterisks in Table 1).

Other aspects of the complex subject of conformity have recently become of interest to researchers studying a variety of taxa, from fish to humans^{26–27}. Here we have shown a non-human species conforming to a group norm, despite possession of an alternative technique that represents the norm of another group. Conformity fits the assumption of an intrinsic motivation to copy others, guided by social bonds rather than material rewards such as food^{5,28}. This has important implications for the transmission and evolution of culture, creating positive feedback that may amplify differences between the traditions of different groups^{11,27}. The substantial variation in the tendency to conform indicated by our results is likely to have additional consequences for the shaping of cultural change.

Our findings also challenge the common assumption that social learning is a function of the juvenile period²⁹. In contrast with data from wild chimpanzees suggesting an early sensitive period in which nut-cracking must be learned¹⁶, and that young males are relatively delayed in the social learning of termite fishing¹⁰, our results show that both sexes can show strong social learning and continue to do so into adulthood.

Finally, it has been argued that evidence for imitation in apes comes largely from individuals that are 'enculturated' by close interactions with humans, shaping their attention in non-natural

ways, whereas other apes may lack the social learning necessary for culture³⁰. Our data, from chimpanzees reared with conspecifics and lacking intense interactions with humans, demonstrate a clear capacity for the cultural transmission of alternative technologies among apes. These results suggest an ancient origin for the conformist cultural propensities so evident in humans.

METHODS

Subjects. Subjects were three separate populations of chimpanzees (*Pan troglodytes*) at the Field Station of the Yerkes Primate Center, median age 15 years (range 6–40 years old). Ages of individual chimpanzees in experimental groups are shown in Fig. 2, and ages and notes on the background of all chimpanzees are provided in the Supplementary Information, together with criteria for model choice. Of the 34 chimpanzees in the two experimental groups, at least 25 were known to be reared by their mothers in captivity; only one was reared in a human home and is thus probably human 'enculturated'³⁰.

Procedures. Details of experimental procedures are given in the Supplementary Information.

Coding. Details of recording and coding methods are given in the Supplementary Information. The principal tool techniques analysed here, Lift versus Poke, could be coded unambiguously because of the tool placement and effects on the Pan-pipe. Interobserver reliability for Poke versus Lift actions, based on coding 10 actions for each of 10 chimpanzees from video records, was 100%, and agreement on success (food delivery) was 94.2%.

Statistics. Owing to the open group context in which the chimpanzees were tested, some chimpanzees performed the task over a thousand times while the experimenter waited for the responses of those chimpanzees that performed least. For this reason, a cap of 30 was set on the number of initial responses by each chimpanzee that would be used to test the hypothesis that alternative traditions had arisen in the two groups at T1. Figure 2a, b summarizes these results. The analysis of conformity required a different approach, because this relied upon individuals recorded as performing both techniques, which some chimpanzees only did after they had achieved 30 successes. Accordingly, for the conformity analysis, total frequencies (tabulated in Supplementary Table S3) were used in the comparison between T1 and T2. Further details of statistical methods used are in the Supplementary Information.

Received 9 June; accepted 21 July 2005.

Published online 21 August 2005.

- Whiten, A. et al. Cultures in chimpanzees. *Nature* **399**, 682–685 (1999).
- Lefebvre, L. in *The Evolution of Cognition* (eds Heyes, C. & Huber, L.) 311–328 (MIT Press, Cambridge, 2000).
- Rendell, L. & Whitehead, H. Cultures in whales and dolphins. *Behav. Brain Sci.* **24**, 309–382 (2001).

4. Whiten, A. *et al.* Charting cultural variation in chimpanzees. *Behaviour* **138**, 1489–1525 (2001).
5. de Waal, F. B. M. *The Ape and the Sushi Master: Cultural Reflections of a Primatologist* (Basic Books, New York, 2001).
6. van Schaik, C. P. *et al.* Orangutan cultures and the evolution of material culture. *Science* **299**, 102–105 (2003).
7. Frigaszy, D. M. & Perry, S. (eds) *The Biology of Traditions: Models and Evidence* (Cambridge Univ. Press, Cambridge, 2003).
8. Whiten, A., Horner, V. & Marshall-Pescini, S. Cultural Panthropology. *Evol. Anthropol.* **12**, 92–105 (2003).
9. McGrew, W. C. *The Cultured Chimpanzee: Reflections on Cultural Primatology* (Cambridge Univ. Press, Cambridge, 2004).
10. Lonsdorf, E. V., Pusey, A. E. & Eberly, L. Sex differences in learning in chimpanzees. *Nature* **428**, 715–716 (2004).
11. Richerson, P. J. & Boyd, R. *Not by Genes Alone: How Culture Transformed Human Evolution* (Chicago Univ. Press, Chicago, 2005).
12. Whiten, A. The second inheritance system of chimpanzees and humans. *Nature* doi:10.1038/nature04023 (in the press).
13. Matsuzawa, T. in *Chimpanzee Cultures* (eds Wrangham, R. W., McGrew, W. C., de Waal, F. B. M. & Heltne, P.) 351–370 (Harvard Univ. Press, Cambridge, 1994).
14. Galef, B. G. Jr. in *The Biology of Traditions: Models and Evidence* (eds Fragaszy, D. & Perry, S.) 159–186 (Cambridge Univ. Press, Cambridge, 2003).
15. Laland, K. N. & Hoppitt, W. Do animals have culture? *Evol. Anthropol.* **12**, 150–159 (2003).
16. Biro, D. *et al.* Cultural innovation and transmission of tool use in wild chimpanzees: evidence from field experiments. *Anim. Cogn.* **6**, 213–223 (2004).
17. Whiten, A., Horner, V., Litchfield, C. A. & Marshall-Pescini, S. How do apes ape? *Learn. Behav.* **32**, 36–52 (2004).
18. Tomasello, M. *The Cultural Origins of Human Cognition* (Harvard Univ. Press, Cambridge, 1999).
19. Menzel, E. W. Jr, Davenport, R. K. & Rogers, C. M. Proto-cultural aspects of chimpanzee's responsiveness to novel objects. *Folia Primatol. (Basel)* **17**, 161–170 (1972).
20. Laland, K. N. & Plotkin, H. C. Social learning and social transmission of digging for buried food in Norway rats. *Anim. Learn. Behav.* **18**, 246–251 (1990).
21. Tonooka, R., Tomonaga, M. & Matsuzawa, T. Acquisition and transmission of tool making and use for drinking juice in a group of captive chimpanzees. *Jap. Psychol. Res.* **39**, 253–265 (1997).
22. Huffman, M. A. & Hirata, S. An experimental study of leaf swallowing in captive chimpanzees: insights into the origin of a self-medicative behaviour and the role of social learning. *Primates* **45**, 113–118 (2004).
23. Asch, S. E. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychol. Monogr.* **70**, 1–70 (1956).
24. Sherif, M. & Murphy, G. *The Psychology of Social Norms* (Harper and Brother, New York, 1936).
25. Jacobs, R. C. & Campbell, D. T. The perpetuation of an arbitrary tradition through several generations of a laboratory microculture. *J. Abnorm. Soc. Psychol.* **62**, 649–658 (1961).
26. Kendal, R. L., Coolen, I. & Laland, K. N. The role of conformity in foraging when personal and social information conflict. *Behav. Ecol.* **15**, 269–277 (2004).
27. Henrich, J. & McElreath, R. The evolution of cultural evolution. *Evol. Anthropol.* **12**, 123–135 (2003).
28. Matsuzawa, T., *et al.* in *Primate Origins of Human Cognition and Behaviour* (ed. Matsuzawa, T.) 557–574 (Springer-Verlag, Tokyo, 2001).
29. Bruner, J. S. Nature and uses of immaturity. *Am. Psychol.* **27**, 687–708 (1972).
30. Call, J. & Tomasello, M. in *Reaching into Thought: The Minds of the Great Apes* (eds Russon, A. E., Bard, K. A. & Parker, S. T.) 371–403 (Cambridge Univ. Press, Cambridge, 1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This project was conducted at the Yerkes National Primate Research Center's Field Station and supported by a project grant from the BBSRC to A.W., by the National Institutes of Health, the Living Links Center of Emory University and the University of St Andrews. A.W. was supported by a Leverhulme Fellowship. We are grateful to M. Dindo, K. Bonnie, J. Rybak and Yerkes Field Station Engineering for logistical support. We also thank Yerkes' animal care and veterinary staff for maintaining the health of the chimpanzees. We are grateful to A. Burnley for construction of the Pan-pipes and to G. Brown, E. Flynn, K. Laland and A. Mesoudi for comments on manuscripts. The Yerkes Center is fully accredited by the American Association for Accreditation for Laboratory Animal Care.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.W. (a.whiten@st-and.ac.uk).

Ca²⁺/calmodulin is critical for brassinosteroid biosynthesis and plant growth

Liqun Du¹ & B. W. Poovaiah¹

Brassinosteroids are plant-specific steroid hormones^{1,2} that have an important role in coupling environmental factors, especially light, with plant growth and development³. How the endogenous brassinosteroids change in response to environmental stimuli is largely unknown. Ca²⁺/calmodulin has an essential role in sensing and transducing environmental stimuli^{4,5}. *Arabidopsis* DWARF1 (DWF1) is responsible for an early step in brassinosteroid biosynthesis that converts 24-methylenecholesterol to campesterol^{6,7}. Here we show that DWF1 is a Ca²⁺/calmodulin-binding protein and this binding is critical for its function. Molecular genetic analysis using site-directed and deletion mutants revealed that loss of calmodulin binding completely abolished the function of DWF1 *in planta*, whereas partial loss of calmodulin binding resulted in a partial dwarf phenotype in complementation studies. These results provide direct proof that Ca²⁺/calmodulin-mediated signalling has a critical role in controlling the function of DWF1. Furthermore, we observed that DWF1 orthologues from other plants have a similar Ca²⁺/calmodulin-binding domain, implying that Ca²⁺/calmodulin regulation of DWF1 and its homologues is common in plants. These results raise the possibility of producing size-engineered crops by altering the Ca²⁺/calmodulin-binding property of their DWF1 orthologues.

The importance of brassinosteroids in plant growth and development is well demonstrated by *Arabidopsis* mutants losing their ability to synthesize or perceive brassinosteroids. These mutants usually show a characteristic pleiotropic phenotype including dwarfism with dark-green colour, photomorphogenesis in the dark, delayed senescence, and reduced apical dominance and fertility^{1,2}. The understanding of the brassinosteroid biosynthetic pathway, which ends in the production of brassinolide, has greatly expedited brassinosteroid signalling research^{8,9}. The endogenous homeostasis of this class of powerful growth regulators is delicately tuned by feedback controls at several steps in the brassinosteroid biosynthetic pathway, such as DWF4, CPD and BR6OX, through the involvement of brassinosteroid perception and connection of the downstream pathway to transcriptional regulation^{8,9}. Recent reports have revealed some exciting results concerning brassinosteroid signal perception and signal relay to the nucleus, and key components of brassinosteroid signalling, such as BRI1 (ref. 10), BAK1 (refs 11, 12), BIN2 (ref. 13) and BES1/BZR1 (refs 1, 2), have been identified and characterized. However, how brassinosteroid action is altered by environmental stimuli remains largely unknown. Recently Pra2, a light-responsive small G protein, was found to interact with and regulate DDWF1, a cytochrome P450 from pea that converts typhasterol to castasterone¹⁴. This is one possible way for light to regulate endogenous brassinosteroid levels. Other environmental signalling events may crosstalk with brassinosteroid signalling and modulate the endogenous brassinosteroid homeostasis; however, little is known about these events⁹.

While screening an *Arabidopsis* complementary DNA expression

library using ³⁵S-calmodulin (CaM) as a probe, a positive clone was isolated that included the 44-amino-acid carboxy terminus of *Arabidopsis* DWF1 (At3g19820). To confirm further the CaM-binding ability of DWF1, the positive clone from the library screening was expressed in *Escherichia coli* and tested for ³⁵S-CaM binding¹⁵ in the presence of Ca²⁺, EGTA, a Ca²⁺ chelator, or Mg²⁺. Recombinant DWF1 bound to ³⁵S-CaM in the presence of Ca²⁺, but not EGTA (Fig. 1a) or Mg²⁺ (data not shown), indicating that DWF1 binds to CaM in a Ca²⁺-dependent manner. C-terminal Flag-tagged DWF1 overexpression lines of *Arabidopsis* were produced to study the DWF1–CaM interaction *in vivo*. Because of the dynamic nature of the interaction between DWF1 and Ca²⁺/CaM, co-immunoprecipitation using a prior *in vivo* crosslinking strategy¹⁶ was adapted to verify their interaction under physiological conditions. The results revealed that Flag-tagged DWF1 interacted with CaM *in vivo* (Fig. 1b). To map its CaM-binding domain (CaMBD), three different segments that covered the entire length of the DWF1 gene were expressed and tested for ³⁵S-CaM binding, and the results indicated that there is only one CaMBD in the 44-amino-acid C terminus (Fig. 1c). The CaMBD is usually comprised of a stretch of 12–30 contiguous amino acids that contain a positive net charge and have a propensity to form an amphiphilic α -helix⁴. Helical wheel projection indicated that the amino acid region 518–539 is the only possible CaMBD (Fig. 1d). A peptide corresponding to 518–539 was synthesized, and its CaM-binding ability was confirmed using PCM6 (ref. 17) in a gel mobility shift assay¹⁵ (Fig. 1e). It was further confirmed to bind to conserved CaM isoforms such as PCM1 (ref. 17) and bovine CaM, but not to *Arabidopsis* CaM9, a divergent CaM isoform (ref. 18 and data not shown). To estimate the affinity of the interaction between DWF1 and Ca²⁺/CaM, fluorescence titration of dansyl-CaM (Sigma) in the presence of increasing amounts of the DWF1 CaM-binding peptide (518–539) was performed as described¹⁹, and the dissociation constant (*K*_d) was calculated to be 24.25 ± 4.62 nM (Fig. 1f).

CaM is a well-characterized Ca²⁺ sensor protein and it can bind to and regulate the function of various CaM-binding proteins by changing their conformation^{4,5}. Overexpression of other brassinosteroid biosynthetic genes such as DWF5 (ref. 20), DET2 (ref. 21), DWF4 (ref. 10) and CPD²² resulted in a hypermorphic phenotype with increased vegetative growth. Surprisingly, overexpression of DWF1 did not show any phenotypic change⁶, implying that an unknown condition was not met for the overexpressed DWF1 to produce a hypermorphic phenotype. In an earlier report we found that upregulated expression of one CaM isoform (PCM1) resulted in increased growth and apical dominance in potato¹⁷, similar to the results of brassinosteroid biosynthetic gene overexpression. The identification of DWF1 as a CaM-binding protein indicates the likelihood of a regulatory mechanism exerted on DWF1 at the protein level.

Because there are no established enzymatic procedures to test the

¹Center for Integrated Biotechnology and Department of Horticulture, Washington State University, Pullman, Washington 99164-6414, USA.

function of DWF1, mutation and complementation strategies were used to study the effect of CaM-binding on DWF1 function. To circumvent the infertility problem of homozygous *dwf1* mutants^{6,7}, a heterozygous *dwf1* line (SALK_006932, ref. 23) was chosen for transformation. A T-DNA insert is shown in Supplementary Fig. 1A, and detailed characterizations confirmed that it is a heterozygous *dwf1* null mutation line (Fig. 2a, b). Heterozygous SALK_006932 plants were transformed with a 35S::DWF1 construct (Supplementary Fig. 1B). Almost all of the T1 plants showed a

normal phenotype. Using polymerase chain reaction (PCR) analysis, one dwarf and seven normal plants were identified that carried at least one copy of the 35S::DWF1 complementary construct and were homozygous knockout at the endogenous DWF1 loci. The dwarfism in the dwarf plant was confirmed to be caused by a failure in expression of the transgene (Fig. 2c, d). The overexpressed (numbers 1, 2, 3, 8) and equally expressed (numbers 5, 7) lines grew like the wild type, confirming a previous report⁶. The underexpressed line (number 4) had an appearance very similar to the wild type with a slightly shorter petiole and rounder leaves, whereas the silenced line (number 6) showed a complete dwarf phenotype (Fig. 2c), indicating that the phenotype of the DWF1 complemented plants (cW) correlated with the transgene expression level in the underexpressed conditions. At the adult stage, 7 out of the 8 complemented lines reached similar levels of height and fertility as that of wild-type plants (Fig. 3c). These results demonstrated that the 35S promoter-driven DWF1 gene can completely rescue the dwarf phenotype of the *dwf1* null mutation.

Three mutants were produced to disrupt the conserved secondary structure of the CaMBD of DWF1 (Supplementary Fig. 1B): V531D (M1), aimed at disturbing the hydrophobic surface;

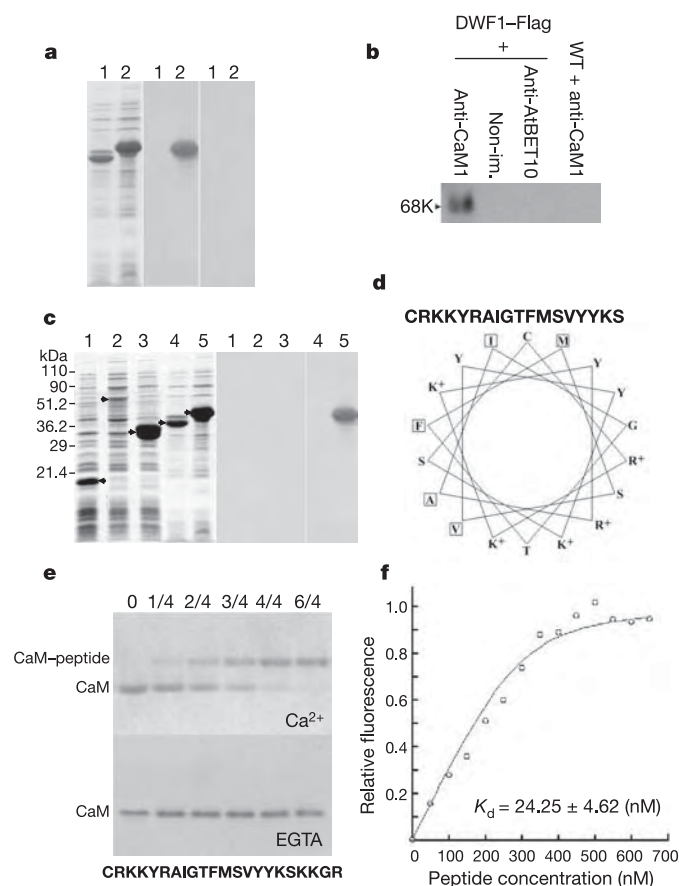


Figure 1 | Ca^{2+} /CaM-binding property of DWF1. **a**, ^{35}S -CaM-binding assay. Total proteins from *E. coli* expressing pScreen-1b (lane 1) or expressing pScreen-1b-517–561 (lane 2) were stained (left panel) or bound to ^{35}S -CaM in the presence of Ca^{2+} (middle panel) or EGTA (right panel), a Ca^{2+} chelator. **b**, Co-immunoprecipitation of Flag-tagged DWF1 and CaM. Total proteins from plants expressing the Flag-tagged DWF1 (DWF1-Flag) or wild-type plants (WT) were incubated with antibodies against full-length human CaM1 (anti-CaM1), AtBET10 (anti-AtBET10) or non-immune serum (Non-im.); immunocomplexes were separated by SDS-PAGE and detected with anti-Flag M2 monoclonal antibody. **c**, Mapping of the CaMBD of DWF1. Total proteins from *E. coli* expressing pET32b (lane 1), pET32b-1–349 (lane 2), pET32b-348–518 (lane 3), pScreen-1b (lane 4) or pScreen-1b-517–561 (lane 5) were stained (left) or were bound to ^{35}S -CaM (right). **d**, Helical wheel projection of amino acids 518–535 from DWF1 CaMBD (518–539). Positively charged residues are marked with a plus sign, whereas hydrophobic residues are boxed; amino acids 518–535 exhibit a typical conserved CaM-binding secondary structure: a typical amphiphilic α -helix with a positively charged face on one side and a predominantly hydrophobic face on the other side. **e**, Gel mobility shift assay of synthetic DWF1 CaMBD: the gel mobility shift was tested using 120 pmol potato CaM6 and increasing amounts of synthetic peptide (peptide/CaM molar ratios are indicated) in the presence of 1 mM Ca^{2+} (top) or 5 mM EGTA (bottom). CaM and the peptide–CaM complex is indicated. **f**, Interaction between dansyl-CaM and the CaM-binding peptide (CRKKYRAIGTFMSVYKSKKGR) of DWF1, as revealed by fluorescence titration.

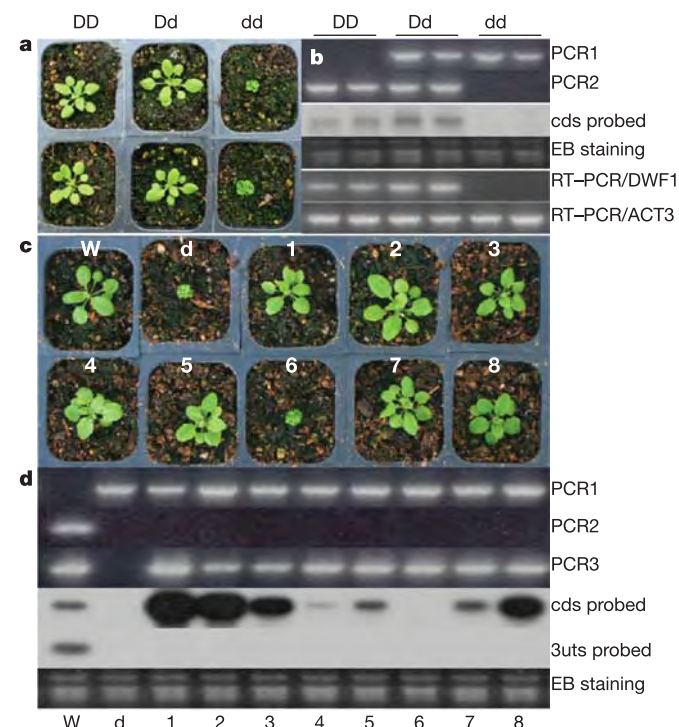


Figure 2 | The DWF1 null mutant line Salk_006932 can be rescued by 35S::DWF1. **a**, Twenty-four-day-old plants (wild type, DD; heterozygous *dwf1* mutant, Dd; homozygous *dwf1* mutant, dd) with Columbia background are shown. **b**, Molecular evidence supporting the genotype of plants in **a**. Top panel: positive PCR1 (Lba-1 and D-2 primers) indicates the existence of insertion; negative PCR2 (D-5u and D-2 primers) indicates double insertions in both copies of endogenous DWF1. See Supplementary Fig. 1 for primer positions. Middle panel: northern blot probed with DWF1 coding sequence (cds); ethidium bromide (EB) staining was used as a loading control. Bottom panel: RT-PCR showing the expression of DWF1 with primer D-5u and D-2; expression of actin 3 (U39480) was used as a control. **c**, Twenty-four-day-old plants: wild type, W; dwarf, d; and eight individual T1 mutant plants complemented with the 35S::DWF1 construct. **d**, Molecular evidence supporting the genotype of plants in **c**. Top panel: PCR1 and PCR2 (see **b**); positive PCR3 (D-1 and D-2 primers) indicates the existence of a complementary 35S::DWF1. Bottom panel: northern blot probed with DWF1 cds (detects both transgenic and endogenous transcription) and 3' end untranslated sequence of DWF1 (3uts; detects endogenous transcription only); EB staining was used as a loading control.

KYR521–523DGD (M2; where K521, Y522 and R523 are changed to D521, G522 and D523), aimed at changing the net surface charges; and Δ 518–539 (M3), a total deletion of the CaMBD. The effects of these mutations were tested, and the results revealed that the CaM-binding ability was markedly decreased for V531D and was

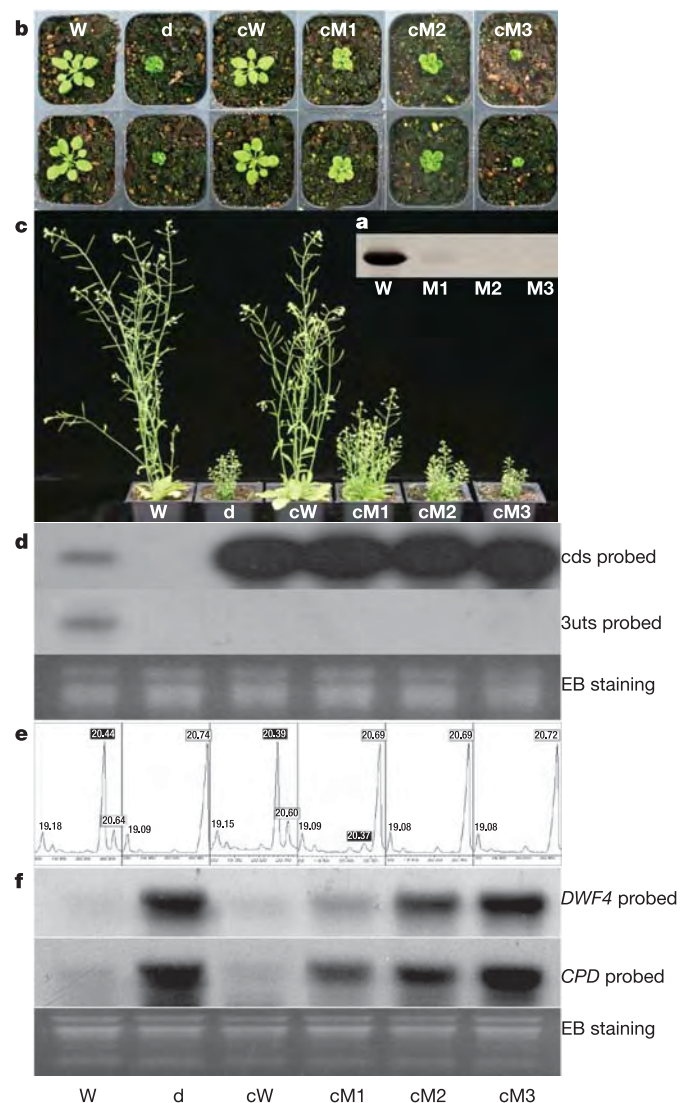


Figure 3 | Loss of CaM binding nullifies DWF1 function. **a**, 35 S-CaM-binding assay of purified recombinant protein (2 μ g per lane) covering amino acid 517 to the stop codon of DWF1(WT), DWF1(V531D), DWF1(KYR521–523DGD) and DWF1(Δ 518–539) (W, M1, M2 and M3, respectively). **b**, Twenty-four-day-old plants: two plants are shown for each line. **c**, Forty-eight-day-old plants of wild type (W), dwarf (d) and dwarf mutants complemented with 35S::DWF1(WT) (cW line 3 in Fig. 2c), DWF1(V531D) (cM1), DWF1(KYR521–523DGD) (cM2) and DWF1(Δ 518–539) (cM3) constructs. **d**, Northern blot hybridized to *cds* and *3uts* probes of DWF1, demonstrating the genotypes of these plants. **e**, GC-MS analysis showing the endogenous 24-methylenecholesterol/campesterol and sitosterol/isofucosterol levels. Sitosterol, isofucosterol and unresolved 24-methylenecholesterol/campesterol peaks are indicated with retention times (min) shown by values in filled boxes, open boxes and without boxes, respectively. The wild-type plants (W) and DWF1 complemented *dwf1* mutant plants (cW) had a very similar sitosterol/isofucosterol pattern with a large sitosterol peak and a very small isofucosterol peak. The *dwf1* mutant (d) as well as cM2 and cM3 had large isofucosterol peaks and no sitosterol peak; cM1 had a large isofucosterol peak and a faint sitosterol peak. **f**, Expression levels of brassinosteroid-regulated DWF4 and CPD (brassinosteroid markers) reflecting endogenous brassinosteroid levels.

undetectable for KYR521–523DGD and Δ 518–539 (Fig. 3a). A similar strategy was used to deliver the mutant *dwf1* complimentary constructs (Supplementary Fig. 1B) into the heterozygous *dwf1* plants. Visual checks identified 18, 11 and 19 dwarf plants in the T1 population carrying 35S::DWF1(V531D), 35S::DWF1(KYR521–523DGD) and 35S::DWF1(Δ 518–539) complimentary constructs, respectively. These plants were confirmed to carry at least one copy of the transgene and the homozygous knockout at the endogenous DWF1 loci. To ensure a reliable comparison of the complementary effects of normal DWF1 and the three mutants, T2 plants with similar expression levels were compared (Fig. 3b–d). In contrast, typical DWF1(V531D) complemented dwarf mutants (cM1) showed some degree of phenotype rescue, with round, expanded leaves but unextended petioles at the rosette stage and bushy flowering stems reaching up to 40% of the height of the wild-type plant. Seeds were produced in most of the slightly shorter siliques of cM1 plants (Fig. 3b, c). Typical DWF1(KYR521–523DGD) complemented plants (cM2) showed very slight phenotypic rescue, with expanded, round and curved leaves and unextended petioles at the rosette stage, as well as bushy and slightly longer flower stems, and seeds in most of the siliques (Fig. 3b, c). All of the DWF1(Δ 518–539) complemented *dwf1* mutants (cM3) grew like the uncomplemented *dwf1* mutants (Fig. 3b, c). The dwarf phenotypes of cM1, cM2, cM3 and the *dwf1* mutant were rescued by exogenous application of brassinolide (data not shown). These results demonstrate that loss-of-function mutations in the CaMBD of DWF1 leads to the functional loss of the whole protein in complementation tests.

DWF1 catalyses the conversion of isofucosterol to sitosterol and 24-methylenecholesterol to campesterol^{6,7}. The endogenous levels of sitosterol/isofucosterol or campesterol/24-methylenecholesterol in wild-type, dwarf and complemented plants were analysed using gas chromatography-mass spectrometry (GC-MS) to monitor the function of the complementing DWF1, DWF1(V531D), DWF1(KYR521–523DGD) and DWF1(Δ 518–539) carried in cW, cM1, cM2 and cM3

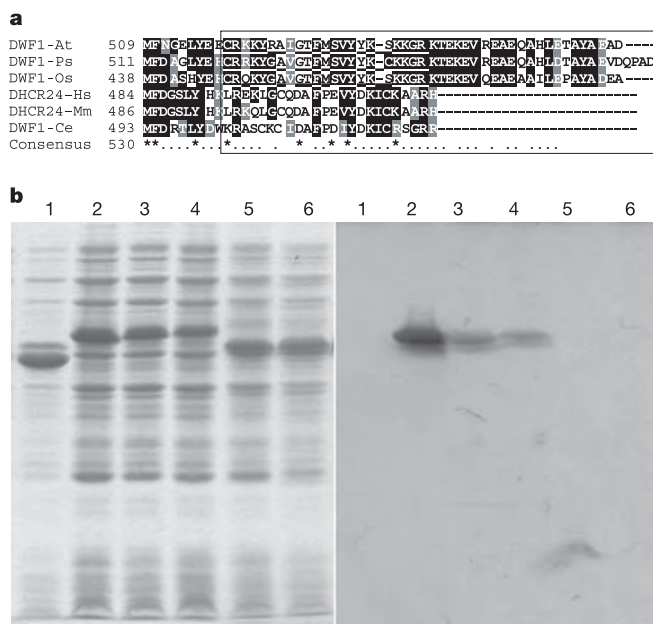


Figure 4 | The C-terminal CaMBD is unique to plant orthologues of DWF1. **a**, Comparison of the C terminus of DWF1 orthologues from *Arabidopsis* (At), pea (Ps), rice (Os), human (Hs), mouse (Mm) and *C. elegans* (Ce). The CaMBD of each plant orthologue is underlined. **b**, 35 S-CaM-binding assay of DWF1 orthologues. Total proteins from *E. coli* expressing pScreen-1b (lane 1) or pScreen-1b carrying the boxed segment in **a** of DWF1 orthologues from *Arabidopsis* (lane 2), pea (3), rice (4), human (5) and mouse (6) were stained (left) or bound to 35 S-CaM in the presence of Ca^{2+} (right).

plants, respectively. The results (Fig. 3e) suggest that the function of the DWF1 enzyme in cW plants was recovered to a level comparable to that in wild-type plants. In cM1 plants DWF1 function was restored to some extent, but the recovery of DWF1 function in cM2 and cM3 plants was almost non-existent. To confirm further that the endogenous brassinosteroid levels are also correlated with complementation events, we monitored the expression of brassinosteroid marker genes such as *CPD* and *DWF4*, which are known to be feedback-controlled by brassinosteroid signals^{8,9}. Our results revealed that in wild-type and cW plants the expression of both *DWF4* and *CPD* was very low, indicating a normal brassinosteroid level; however, in *dwf1* mutant and cM1, cM2 and cM3 plants expression of both *DWF4* and *CPD* was upregulated, indicating decreased levels of endogenous brassinosteroids (Fig. 3f). Furthermore, transcription of *DWF4* and *CPD* correlated with the phenotypes of the corresponding plants and their endogenous sitosterol/isofucosterol levels, as revealed by GC-MS analysis.

DWF1 orthologues exist in both plants and animals. Protein sequences of DWF1 and its orthologues from pea, rice, human, mouse and *Caenorhabditis elegans* were aligned and the C-terminal CaMBD was found to be conserved in plants but not animals (Fig. 4a). The boxed parts in Fig. 4a—the comparable parts of DWF1 and its orthologues from *Arabidopsis*, pea, rice, human and mouse—were expressed in *E. coli*, and the ³⁵S-CaM-binding assay demonstrated that only the C terminus of plant DWF1 orthologues bind to Ca²⁺/CaM, indicating that the orthologues of DWF1 in plants are probably regulated by Ca²⁺/CaM through the C-terminal region. Whether portions other than the C terminus of the DWF1 orthologues from animals bind to Ca²⁺/CaM is beyond the focus of this research.

Characterization of brassinosteroid synthetic mutants such as *det2* (ref. 24) and *cpd*²² linked the light signal to brassinosteroids, and this connection has been established by genetic and biochemical analyses of the brassinosteroid signalling processes^{2,3}. Light has been shown to induce cytosolic Ca²⁺ changes and subsequent physiological responses related to photomorphogenesis, and probably phototropism^{25,26}, indicating that light signal can be transduced through Ca²⁺. The *Arabidopsis det3* mutant expresses remarkably reduced subunit C of the vacuolar H⁺-ATPase (VHA-C) due to a mutation in this gene, and its de-etiolated phenotype and dwarfism indicate a defect in hormone signalling²⁷. The *det3* mutant was later found to have altered Ca²⁺ oscillation and guard cell response to some of the tested signals, probably caused by the deficient VHA-C activity²⁸. These results support the hypothesis that a defective growth phenotype and possibly altered hormone signalling in *det3* is connected to altered Ca²⁺ responses.

Ca²⁺ is an important second messenger and an integral part of the cellular signalling system in plants^{5,29}. Recently, we were prompted to survey the CaM-binding properties of other proteins involved in brassinosteroid biosynthesis. Interestingly, two important enzymes, DWF4 and CPD, were also found to be Ca²⁺/CaM-binding proteins, especially DWF4, which has two CaM-binding domains with markedly different affinities for Ca²⁺/CaM (our own unpublished data). The results described herein suggest that Ca²⁺/CaM binding is critical for DWARF1 function in brassinosteroid biosynthesis and plant growth, and that Ca²⁺/CaM may have complex and wide-ranging regulatory controls on the biosynthesis of brassinosteroid; DWF1 seems to be one of these regulatory steps.

METHODS

Co-immunoprecipitation with prior crosslinking. One gram of rosette leaves from 3–4-week-old *Arabidopsis* plants expressing DWF1-Flag were harvested, washed briefly, cut into ~5-mm² pieces, soaked in 40 ml ice-cold lysis buffer (10 mM HEPES-KOH, pH 7.4, 12.5% sucrose, 10 mM KAC, 3 mM MgCl₂, 1 mM CaCl₂) supplemented with 0.01% detergent Silwet-77 and 1% formaldehyde, and subjected to vacuum and release for two cycles (5 min each). After incubating at 4 °C for 30 min, the leaves were washed two times for 20 min

with 50 ml lysis buffer supplemented with 0.3 M glycine at 4 °C, and stored at –80 °C until further use. The fixed leaves were ground in liquid N₂ to a fine powder, resuspended in four volumes of lysis buffer supplemented with Na₃N₃ and 1 × Roche complete protease inhibitors, and centrifuged two times at 4 °C, 12,000 g for 10 min to obtain a clear total protein extract. Polyclonal anti-CaM1 antibody (2 µg, Santa Cruz) was added to 500 µg total protein and the mixture was shaken at room temperature for 1 h. Pre-blocked protein-G plus agarose was added and shaken for another hour. The immunocomplexes were washed three times in lysis buffer supplemented with 1 × Roche complete protease inhibitors and 150 mM NaCl, resuspended in 30 µl 2 × SDS loading buffer, boiled for 25 min to reverse the crosslinking induced by formaldehyde, separated on SDS-PAGE, and immuno-detected with anti-Flag M2 monoclonal antibody. Total protein from wild-type plants, non-immune serum and anti-AtBET10 (ref. 15) antibody were used as negative controls.

Sterol analysis. Six- to seven-week-old greenhouse-grown plants were harvested and ground into a fine powder in liquid nitrogen. Aliquots of 100 mg ground material were saponified in 1 N NaOH, 70% ethanol at 70 °C for 2 h. The sterols were extracted three times with 500 µl hexane, and the combined extracts were dried and dissolved in 100 µl of methylene chloride. GC-MS analyses were conducted on an Agilent 6890 GC system with 5973N mass selective detector. A 1 µl aliquot of methylene chloride extract containing the underivatized sterols was loaded by cool on column injection on a Zebron (Phenomenex) ZB-5 column, 30 m × 0.25 mm internal diameter with 0.25 µm film thickness. Chromatography was accomplished using He as the carrier gas at a flow rate of 0.7 ml min^{–1}. The column oven temperature was programmed to run from 40 °C to 300 °C increased at 20 °C min^{–1}, with a hold for 10 min at 300 °C. Identification and quantification of sterols was accomplished by comparison to authentic standards based on the retention time, peak area and mass fragmentation pattern.

Received 11 January; accepted 1 July 2005.

- Clouse, S. D. Brassinosteroid signal transduction: clarifying the pathway from ligand perception to gene expression. *Mol. Cell* **10**, 973–982 (2002).
- Nemhauser, J. L. & Chory, J. Bring it on: new insights into the mechanism of brassinosteroid action. *J. Exp. Bot.* **55**, 265–270 (2004).
- Clouse, S. D. Integration of light and brassinosteroid signals in etiolated seedling growth. *Trends Plant Sci.* **6**, 443–445 (2001).
- Snedden, W. A. & Fromm, H. Calmodulin as a versatile calcium signal transducer in plants. *New Phytol.* **151**, 35–66 (2001).
- Yang, T. & Poovaiah, B. W. Calcium/calmodulin-mediated signal network in plants. *Trends Plant Sci.* **8**, 505–512 (2003).
- Klahre, U. et al. The *Arabidopsis* *DIMINUTO/DWARF1* gene encodes a protein involved in steroid synthesis. *Plant Cell* **10**, 1677–1690 (1998).
- Choe, S. et al. The *Arabidopsis* *dwarf1* mutant is defective in the conversion of 24-methylenecholesterol to campesterol in brassinosteroid biosynthesis. *Plant Physiol.* **119**, 897–907 (1999).
- Fujioka, S. & Yokota, T. Biosynthesis and metabolism of brassinosteroids. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **54**, 137–164 (2003).
- Thummel, C. S. & Chory, J. Steroid signalling in plants and insects—common themes, different pathways. *Genes Dev.* **16**, 3113–3129 (2002).
- Wang, Z. Y., Seto, H., Fujioka, S., Yoshida, S. & Chory, J. BRI1 is a critical component of a plasma-membrane receptor for plant steroids. *Nature* **410**, 380–383 (2001).
- Li, J. et al. BAK1, an *Arabidopsis* LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signalling. *Cell* **110**, 213–222 (2002).
- Nam, K. H. & Li, J. BRI1/BAK1, a receptor kinase pair mediating brassinosteroid signalling. *Cell* **110**, 203–212 (2002).
- Li, J. & Nam, K. H. Regulation of brassinosteroid signalling by a GSK3/SHAGGY-like kinase. *Science* **295**, 1299–1301 (2002).
- Kang, J. G. et al. Light and brassinosteroid signals are integrated via a dark-induced small G protein in etiolated seedling growth. *Cell* **105**, 625–636 (2001).
- Du, L. & Poovaiah, B. W. A novel family of Ca²⁺/calmodulin-binding proteins involved in transcriptional regulation: Interaction with fsh/Ring3 class transcription activators. *Plant Mol. Biol.* **54**, 549–569 (2004).
- Hall, D. B. & Struhl, K. The VP16 activation domain interacts with multiple transcriptional components as determined by protein-protein cross-linking in vivo. *J. Biol. Chem.* **277**, 46043–46050 (2002).
- Poovaiah, B. W., Takezawa, D., An, G. & Han, T. J. Regulated expression of a calmodulin isoform alters growth and development in potato. *J. Plant Physiol.* **149**, 553–558 (1996).
- Zielinski, R. E. Characterization of three new members of the *Arabidopsis thaliana* calmodulin gene family: conserved and highly diverged members of the gene family functionally complement a yeast calmodulin null. *Planta* **214**, 446–455 (2002).
- Bouche, N., Scharlat, A., Snedden, W., Bouchez, D. & Fromm, H. A novel family of calmodulin-binding transcription activators in multicellular organisms. *J. Biol. Chem.* **277**, 21851–21861 (2002).

20. Choe, S. *et al.* Lesions in the sterol Δ^7 reductase gene of *Arabidopsis* cause dwarfism due to a block in brassinosteroid biosynthesis. *Plant J.* **21**, 431–443 (2000).
21. Chory, J. & Li, J. Gibberellins, brassinosteroids and light-regulated development. *Plant. Cell Environ.* **20**, 801–806 (1997).
22. Szekeres, M. *et al.* Brassinosteroids rescue the deficiency of CYP90, a cytochrome P450, controlling cell elongation and de-etiolation in *Arabidopsis*. *Cell* **85**, 171–182 (1996).
23. Alonso, J. M. *et al.* Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301**, 653–657 (2003).
24. Li, J., Nagpal, P., Vitart, V., McMorris, T. C. & Chory, J. A role for brassinosteroids in light-dependent development of *Arabidopsis*. *Science* **272**, 398–401 (1996).
25. Shacklock, P. S., Read, N. D. & Trewavas, A. J. Cytosolic free calcium mediates red light-induced photomorphogenesis. *Nature* **358**, 753–755 (1992).
26. Baum, G., Long, J. C., Jenkins, G. I. & Trewavas, A. J. Stimulation of the blue light phototropic receptor NPH1 causes a transient increase in cytosolic Ca^{2+} . *Proc. Natl Acad. Sci. USA* **96**, 13554–13559 (1999).
27. Schumacher, K. *et al.* The *Arabidopsis det3* mutant reveals a central role for the vacuolar H^{+} -ATPase in plant growth and development. *Genes Dev.* **13**, 3259–3270 (1999).
28. Allen, G. J. *et al.* Alteration of stimulus-specific guard cell calcium oscillations

and stomatal closing in *Arabidopsis det3* mutant. *Science* **289**, 2338–2342 (2000).

29. Poovaiah, B. W. & Reddy, A. S. Calcium messenger system in plants. *CRC Crit. Rev. Plant Sci.* **6**, 47–103 (1987).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was partially supported by grants from the National Science Foundation and the United States Department of Agriculture. We thank T. Yang and other members of the laboratory for help and advice during the screening of the cDNA expression library and for discussions during manuscript preparation; M. Zhu for help during fluorescence spectrum analysis; and R. Ketchum and S. Mattinson for help in GC-MS analysis and interpretation. We appreciate the critical reading and helpful suggestions of D. von Wettstein and D. Takezawa.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.W.P. (poovaiah@wsu.edu).

LETTERS

WntD is a feedback inhibitor of Dorsal/NF- κ B in *Drosophila* development and immunity

Michael D. Gordon¹, Marc S. Dionne², David S. Schneider² & Roel Nusse¹

Regulating the nuclear factor- κ B (NF- κ B) family of transcription factors is of critical importance to animals, with consequences of misregulation that include cancer, chronic inflammatory diseases and developmental defects¹. Studies in *Drosophila melanogaster* have proved fruitful in determining the signals used to control NF- κ B proteins, beginning with the discovery that the Toll/NF- κ B pathway, in addition to patterning the dorsal–ventral axis of the fly embryo, defines a major component of the innate immune response in both *Drosophila* and mammals^{2,3}. Here, we characterize the *Drosophila* *wntD* (Wnt inhibitor of Dorsal) gene. We show that WntD acts as a feedback inhibitor of the NF- κ B homologue Dorsal during both embryonic patterning and the innate immune response to infection. *wntD* expression is under the control of Toll/Dorsal signalling, and increased levels of WntD block Dorsal nuclear accumulation, even in the absence of the I κ B homologue Cactus. The WntD signal is independent of the common Wnt signalling component Armadillo (β -catenin). By engineering a gene knockout, we show that *wntD* loss-of-function mutants have immune defects and exhibit increased levels of Toll/Dorsal signalling. Furthermore, the *wntD* mutant phenotype is suppressed by loss of zygotic *dorsal*. These results describe the first secreted feedback antagonist of Toll signalling, and demonstrate a novel Wnt activity in the fly.

The dorsal–ventral (D–V) axis of the *Drosophila* embryo is initially patterned by a ventral-to-dorsal nuclear gradient of Dorsal protein activity under the control of Spatzle/Toll signalling^{4,5}. Toll activates Dorsal primarily through the degradation of Cactus, thereby freeing Dorsal to enter the nucleus and activate or repress target genes⁶. The transcriptional profile that is regulated by Dorsal defines the spatial organization of tissues in the embryo, with ventral-most cells becoming mesoderm, flanked by the mesectoderm and neuroectoderm in more lateral regions, and gut primordia at the poles⁷.

The gene *wntD* was identified as a member of the *Drosophila* Wnt family based on a genomic search for Wnt-related genes (synonyms CG8458 and *wnt8*⁸). Examination of *wntD* RNA *in situ* revealed that the first detectable expression is seen at the ventral poles of the blastoderm embryo, followed by sequential ventral-to-dorsal expression in the presumptive mesoderm, mesectoderm and neuroectoderm (Fig. 1). Embryos derived from mothers carrying a dominant, activated allele of Toll, express *wntD* RNA more broadly and at higher levels than wild type (Fig. 1c, d). This demonstrates that *wntD* expression is induced by Toll signalling. Examination of WntD protein distribution shows that WntD is secreted and travels multiple cell diameters away from producing cells, suggesting that WntD is capable of signalling at a distance (Fig. 1f, g).

To uncover the signalling activity of WntD in the embryo, we expressed WntD ectopically in the female germline, producing blastoderm stage embryos that contain high levels of WntD protein (data not shown). These embryos lacked detectable nuclear Dorsal

(Fig. 2b); although total cellular levels of Dorsal protein remained unchanged (Fig. 2b, inset). Consequently, the mesodermal Dorsal target gene *Twist* was not expressed (Fig. 2d), and the embryos produced only dorsal cuticle (Fig. 2f). Furthermore, the observed defects were specific to dorsal–ventral patterning, as the anterior–posterior patterning gene *hunchback* was unaffected by WntD (data not shown). After submission of this manuscript, similar results for the overexpression of WntD were reported⁹.

In order to determine the point of intersection between WntD activity and the Toll/Dorsal pathway, we constructed flies that overexpress WntD and carry strong hypomorphic alleles of *cactus*. Although maternal *cactus* mutants exhibited a ventralized phenotype (Fig. 2g), those also overexpressing WntD were dorsalized, and indistinguishable from embryos overexpressing WntD alone (Fig. 2h). These data demonstrate that WntD (a secreted growth

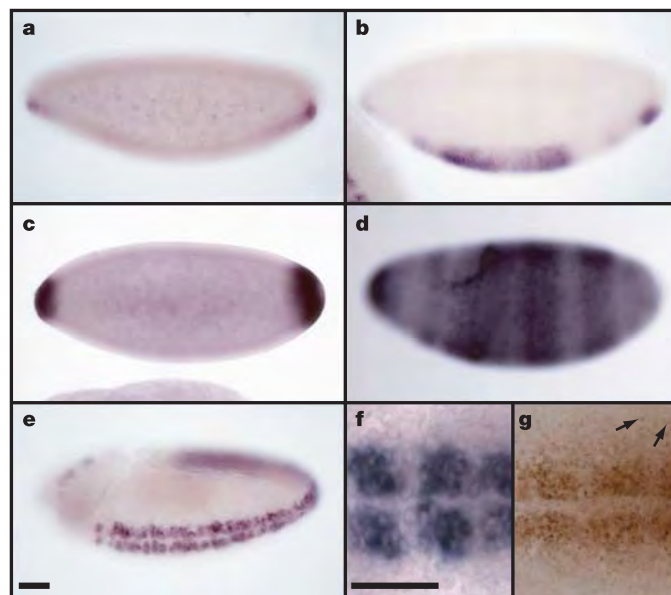


Figure 1 | *wntD* is expressed with D–V polarity, and is under the control of Toll signalling. **a–e**, *In situ* hybridization to *wntD* mRNA in wild-type embryos (**a**, **b** and **e**) and those derived from *Toll*^{10b} mothers (**c** and **d**). Wild type expression is seen in the ventral poles at stage 5 (**a**), presumptive mesoderm at stage 6 (**b**) and neurogenic ectoderm at stage 9 (**e**). Stronger, expanded *wntD* expression is seen in *Toll*^{10b}-derived embryos at stages 5 and 6 (**c** and **d**). **f**, **g**, Close-up ventral views (anterior left) of stage-9 wild-type embryos stained for *wntD* mRNA (**f**) or with anti-WntD antibody (**g**). Arrows indicate examples of WntD antigen detected multiple cell diameters away from *wntD*-expressing cells. Scale bars, 50 μ m. All embryos here and in Figs 2 and 3 oriented anterior left, ventral down, unless otherwise indicated.

¹Department of Developmental Biology, Howard Hughes Medical Institute, Beckman Center and ²Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, California 94305, USA.

factor) is capable of producing a signal that blocks Dorsal nuclear translocation downstream of, or in parallel to, Cactus. It has been shown previously that Dorsal undergoes Toll-dependent and -independent phosphorylation¹⁰, and that Dorsal nuclear localization can be regulated independently of Cactus¹¹.

That WntD is a member of the Wnt family of growth factors raises the question of whether it signals through the well-characterized Frizzled/Armadillo (β -catenin) pathway¹². We suggest that it does not, based on two lines of evidence: (1) germline clones of *axin*, a negative regulator of Armadillo, do not produce dorsalized embryos¹³; and (2) overexpression of WntD in tissues sensitive to Armadillo signalling does not have any detectable effect (data not shown). These observations however, do not rule out the possibility that WntD signals through a Frizzled receptor in an Armadillo-independent manner.

In order to investigate the role of endogenous WntD, we constructed a loss-of-function mutation using 'ends-out' gene targeting (Fig. 3a). The modified *wntD* locus produced no detectable protein, as assayed by western blot (Fig. 3c). Analysis of flies homozygous for either of two knockout alleles (labelled *wntD*^{KO1} or *wntD*^{KO2}) revealed that *wntD* is not essential for viability or fertility.

Despite their viability, *wntD* mutant embryos show an expansion of nuclear Dorsal into the pole regions where endogenous WntD is first detected (Fig. 3e). This indicates that the earliest role of WntD in the embryo is to restrict the field of Dorsal activation, thereby ensuring the establishment of the proper boundary between the developing ventral and terminal domains; Dorsal, along with A-P positional information, induces transcription of *wntD* at the ventral poles of the embryo, and WntD in turn feeds back to repress Dorsal nuclear translocation, and prevent improper spread of the ventral domain. This mechanism stands in contrast to another characterized

mode of Dorsal pathway repression at the embryonic termini—that of signalling from the Torso (Tor) receptor tyrosine kinase¹⁴. In the case of Torso, signalling at the poles of the embryo selectively interferes with the ability of Dorsal to repress the expression of specific target genes, although exerting only a minor effect on those genes activated by Dorsal¹⁴. These data suggest that Torso signalling affects the activity of nuclear Dorsal, whereas WntD signalling affects Dorsal's nuclear translocation.

In addition to its role in D-V patterning, it has been well established that Toll/NF- κ B signalling has a more evolutionarily conserved role in regulating the innate immune system^{3,15}. During the immune response Toll induces the nuclear translocation of two NF- κ B family members: Dorsal and Dorsal-related immunity factor (Dif). Genetic analysis has suggested that Dif, although dispensable for development, is the major transcription factor involved in the Toll-mediated immune response¹⁶. In addition to Dorsal and Dif, the fly immune response also uses a third NF- κ B related protein, Relish, which is activated on signalling by PGRP-LC and Imd^{17,18}. Together, these pathways regulate the expression of hundreds of genes after microbial infection¹⁹.

In light of the interaction between WntD and Dorsal in the embryo, we asked if WntD could have a role later in the fly's life as a repressor of Toll/Dorsal-mediated immunity. Polymerase chain reaction with reverse transcription (RT-PCR) was used to confirm

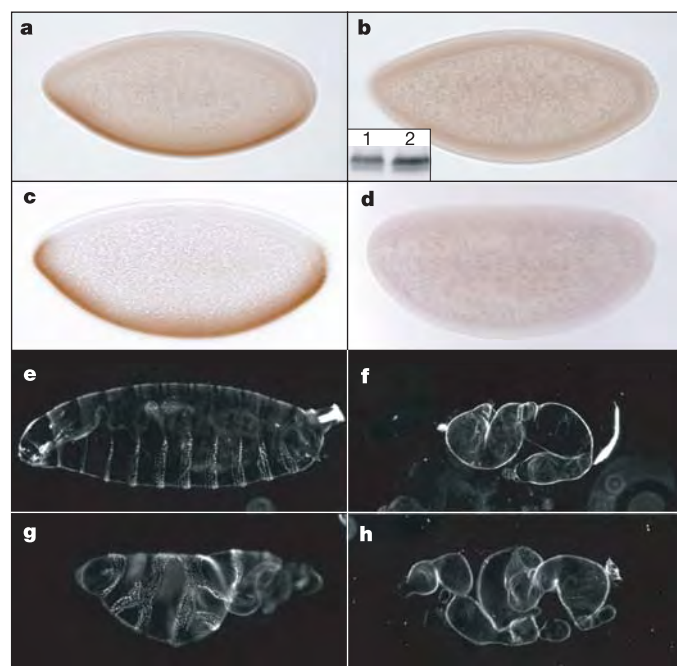


Figure 2 | Overexpression of WntD blocks Dorsal protein activation independently of Cactus. **a, c**, Wild-type embryos stained with antibodies to Dorsal (**a**) or Twist (**c**). **b, d**, Embryos from females carrying *P[nos-Gal4:VP16]* and *P[UASp-wntD]* transgenes, stained with antibodies against Dorsal (**b**) or Twist (**d**). Inset (**b**), total Dorsal protein levels (assayed by western blot) are equivalent in wild-type embryos (lane 1) and those from *P[nos-Gal4:VP16]/P[UASp-wntD]* females (lane 2). **e–h**, Cuticles of embryos with the maternal genotypes: wild type (**e**); *P[nos-Gal4:VP16]/P[UASp-wntD]* (**f**); *cact¹/cact⁴* (**g**); and *cact¹/cact⁴; P[nos-Gal4:VP16]/P[UASp-wntD]* (**h**).

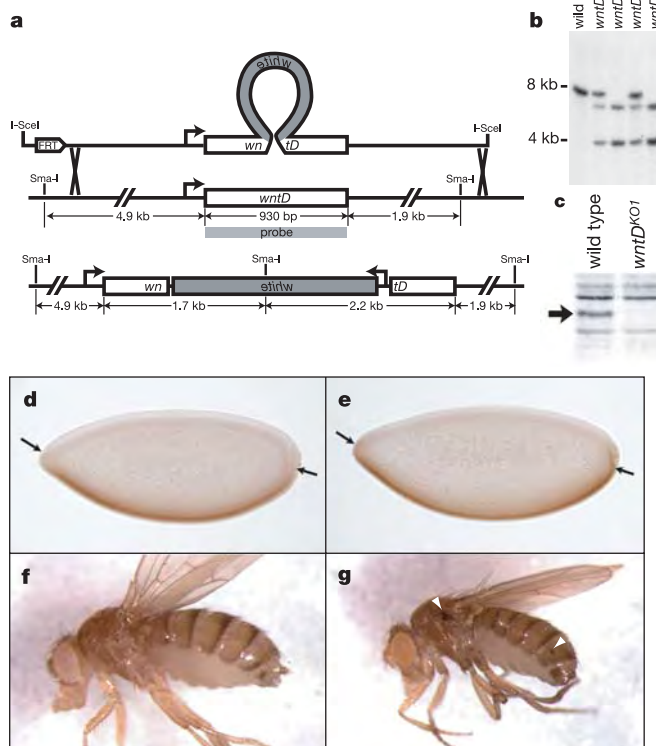


Figure 3 | *wntD* knockout flies exhibit ectopic Dorsal activation. **a**, Ends-out knockout targeting scheme, illustrating how a *white* mini-gene was used to interrupt the *wntD* open reading frame. **b**, Southern blot of *Sma*-I digested genomic DNA, confirming proper integration of targeting construct. TM6 indicates the third chromosome balancer which carries a wild-type *wntD* locus. **c**, Anti-WntD western blot of lysate from wild type and *wntD*^{KO1} embryos (arrow indicates size of WntD protein). **d, e**, Embryos of the genotype *yw* (**d**) and *yw; wntD^{KO1}* (**e**) stained with antibodies against Dorsal. Arrows show point of ventral-most nuclear Dorsal seen in control embryos. **f, g**, Adult female *yw* (**f**) and *yw; wntD^{KO1}* (**g**) flies. White arrowheads mark sites of ectopic melanization.

expression of endogenous *wntD* RNA in adults (data not shown). *wntD* mutant adults appear normal, with the exception that at low frequency (1–2%) we have observed sites of ectopic melanization, most notably on the wing hinge (Fig. 3g). This is consistent with a role for WntD in maintaining low basal levels of Toll/Dorsal signalling, as other mutations that hyper-activate Toll show increased levels of phenoloxidase-driven melanization^{20,21}. Furthermore, Dorsal has been shown to be an essential component of the melanization response in larvae²².

To investigate the role of WntD after septic injury, *wntD* and control flies were injected with a dilute culture of the gram-positive bacterium *Micrococcus luteus*, and the induction of antimicrobial peptide (AMP) transcripts were monitored over time using quantitative RT-PCR (Fig. 4b, c). We observed that some, but not all, AMPs showed aberrant expression in *wntD* mutants. The AMP *dipteriscin* was most severely affected, with *wntD* flies displaying dramatically elevated basal levels of expression (approximately 15-fold), and significantly higher mRNA levels following infection (Fig. 4b). In contrast, *drosomycin* mRNA levels were not significantly different from controls in either uninfected or infected *wntD* mutants. A third AMP, *defensin*, showed an intermediate pattern of expression, with elevated mRNA levels in *wntD* mutants at some time points (data not shown).

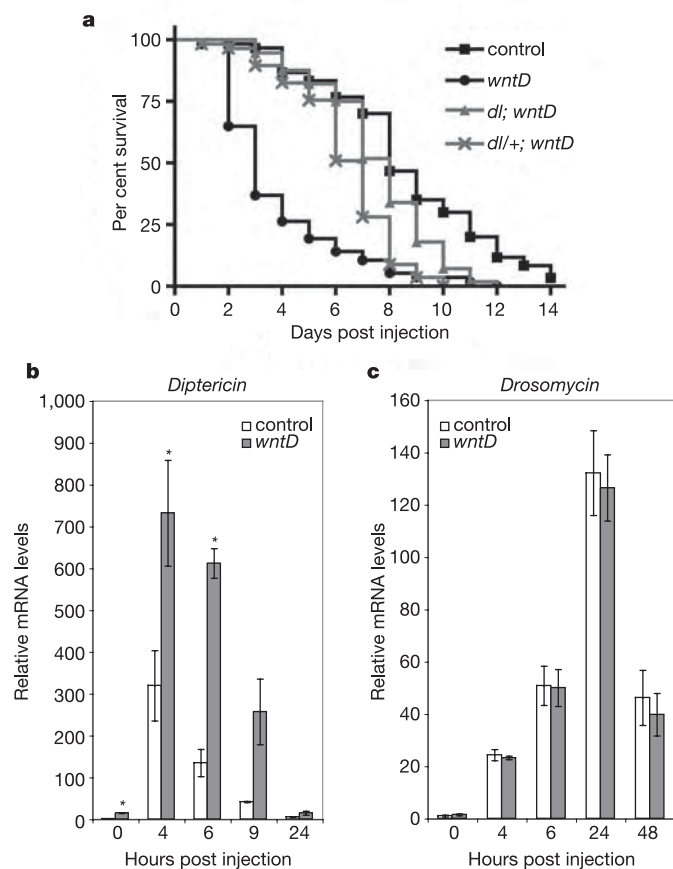


Figure 4 | *wntD* mutants show an aberrant response to microbial infection. **a**, One-week-old adult *yw* (filled squares, $n = 60$), *yw; wntD^{KO1}* (filled circles, $n = 57$), *yw; dl¹/dl⁴; wntD^{KO1}* (grey triangles, $n = 56$) and *yw; dl⁴/+; wntD^{KO1}* (grey crosses, $n = 57$) were injected with a dilute culture of *Listeria monocytogenes*, and survival was monitored. Log rank tests indicate that *wntD* mutant curve is significantly different from the other three, with $P < 0.0001$. **b**, **c**, Real-time PCR was used to monitor *dipteriscin* (**b**) and *drosomycin* (**c**) mRNA levels in *yw* (white bars) and *yw; wntD^{KO1}* (grey bars) adults after injection with *Micrococcus luteus*. Results are mean $\pm$ s.e.m. Asterisks denote significance level of $P < 0.05$.

These results pose an apparent paradox, as previous experiments have characterized *dipteriscin* as a target of IMD/Relish, and *drosomycin* as a target of Toll signalling^{15,23}. *Drosomycin* expression is reported to be primarily regulated by Dif in adult flies, and appears to be unaffected by increased Dorsal activity¹⁶. Thus, our results for *Drosomycin* are consistent with past work. The *dipteriscin* result initially appears puzzling, but existing data demonstrate that the signal transduction pathways regulating immunity are not as specific as initially described. For example, Relish is required for *dipteriscin* induction in response to infections *in vivo*, but constitutive activation of Toll signalling results in elevated levels of *dipteriscin* in adult flies¹⁹. Furthermore, Dorsal is sufficient to activate the *dipteriscin* promoter *in vitro*²⁴. The simplest explanation for these observations is that *dipteriscin* transcription can be induced by Toll/Dorsal signalling. Taken together, these data support a model in which WntD signalling specifically represses Toll/Dorsal, and not Toll/Dif signalling.

Given a role for WntD in the regulation of antimicrobial gene transcription, we sought to determine whether *wntD* mutants were immunocompromised. To test this, *wntD* and control adults were infected with the gram-positive, lethal pathogen *Listeria monocytogenes*²⁵. In response to infection, *wntD* mutants exhibited significantly higher levels of mortality when compared with parental lines (Fig. 4a). Importantly, this phenotype was suppressed by the introduction of *dorsal* mutations (Fig. 4a), with close to full suppression in the absence of both copies of *dorsal* and partial suppression in flies heterozygous for a *dorsal* mutation. These genetic interactions are consistent with our assertion that WntD specifically regulates Dorsal, and not other mediators of immunity. Recent reports have demonstrated that a fly's response to bacterial challenge includes factors that are damaging to the host²⁶, and that increased Toll signalling can render flies more susceptible to viral infection²⁷. We therefore propose that it is the deleterious hyper-activation of specific Dorsal target genes that is responsible for the increased mortality seen in *wntD* mutants. Furthermore, the susceptibility of *wntD* mutants to a lethal infection suggests a reason for the positive selection of *wntD* during evolution; immune responses have a cost, and their appropriate downregulation would be expected to provide flies with a selective advantage. Although *wntD* flies appear healthy in a lab environment, it is easy to imagine that under the more stressful, and septic, conditions in the wild, flies lacking *wntD* would suffer the perils of a hyperactive immune system.

We have presented evidence that WntD, a Wnt family member, produces a signal that blocks the nuclear translocation of Dorsal. Furthermore, WntD is a target of Toll/Dorsal signalling, and creates a negative feedback loop to repress Dorsal activation. We have shown that *wntD* is not required for viability under lab conditions, but that *wntD* mutants show defects in embryonic Dorsal regulation, and in the adult innate immune system. As the WntD signal in the embryo is not mediated by Armadillo, we suppose that the immune function of WntD is also Armadillo-independent, although Zettervall *et al.* have observed immune defects in flies expressing a dominant-negative form of the Armadillo partner DTCE²⁸. Further characterization of signalling events bridging WntD and Dorsal could yield valuable insight into the regulation of the therapeutically important NF- κ B family of proteins.

METHODS

For detailed methods see Supplementary Information.

***Drosophila* stocks.** Flies carrying a *P[UASp-wntD]* were generated using standard techniques. Other transgenic flies used were: *P[nos-Gal4:VP16]* (ref. 29). Knockout methods are described in Supplementary Information.

Immunostaining and *in situ* hybridization. All *in situ* hybridizations were performed using standard procedures, with the exception that Proteinase K treatment was omitted. All immunostainings and cuticle preparations were performed using standard procedures. Primary antibodies were used at the following dilutions: mouse anti-Dorsal (Developmental studies hybridoma bank; 1:10), rabbit anti-Twist (a gift from S. Roth; 1:5,000) and rabbit anti-WntD (1:500). Photos were taken using a Zeiss Axiocam camera, images were

processed with Adobe Photoshop and figures were prepared using Adobe Illustrator.

Western and Southern blots. The western blot for Dorsal protein was performed as previously described¹⁰. Lysates were prepared in TNT buffer (150 mM NaCl, 50 mM Tris-HCl at pH 7.5 and 1% Triton X-100). In order to detect WntD in embryonic lysates, a collection of 0–3 h embryos were lysed in TNT buffer, and incubated overnight with 20 ml Blue Sepharose beads (Amersham Biosciences). The beads were washed, and exposed to sample buffer (15 mM Tris-HCl, 2.5% glycerol, 5% SDS, 1% 2- β -mercaptoethanol and 0.006% bromophenol blue) before gel loading. Western blotting was performed using standard techniques with rabbit anti-WntD antibody (1:1,000). Southern blots were performed using standard techniques. *wntD* radio-labelled probe was generated to full length *wntD* complementary DNA using Rediprime II kit (Amersham Biosciences).

Bacterial injections and quantitative RT-PCR. All injections were done using male flies aged one week post eclosion. A culture of *Listeria monocytogenes* was diluted to an A₆₀₀ of 0.1, and 25 nl was injected abdominally using a pulled glass needle as previously described³⁰. Groups of 20 flies of each genotype were injected in an alternating manner to control for variability over time. Flies were maintained on non-yeasted, standard dextrose medium at 25 °C, 65% relative humidity and survival was monitored daily. *Micrococcus luteus* was injected as described for *L. monocytogenes*. RNA extraction and quantitative RT-PCR was performed as described³⁰ with the exception that six flies were used per sample.

Received 5 June; accepted 20 July 2005.

Published online 17 August 2005.

- Hayden, M. S. & Ghosh, S. Signaling to NF- κ B. *Genes Dev.* **18**, 2195–2224 (2004).
- Beutler, B. Inferences, questions and possibilities in Toll-like receptor signalling. *Nature* **430**, 257–263 (2004).
- Lemaitre, B. The road to Toll. *Nature Rev. Immunol.* **4**, 521–527 (2004).
- Rushlow, C. A., Han, K., Manley, J. L. & Levine, M. The graded distribution of the dorsal morphogen is initiated by selective nuclear transport in *Drosophila*. *Cell* **59**, 1165–1177 (1989).
- Morisato, D. & Anderson, K. V. The *spatzle* gene encodes a component of the extracellular signalling pathway establishing the dorsal-ventral pattern of the *Drosophila* embryo. *Cell* **76**, 677–688 (1994).
- Belvin, M. P., Jin, Y. & Anderson, K. V. Cactus protein degradation mediates *Drosophila* dorsal-ventral signalling. *Genes Dev.* **9**, 783–793 (1995).
- Stathopoulos, A., Van Drenth, M., Erives, A., Markstein, M. & Levine, M. Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo. *Cell* **111**, 687–701 (2002).
- Llimargas, M. & Lawrence, P. A. Seven Wnt homologues in *Drosophila*: a case study of the developing tracheae. *Proc. Natl Acad. Sci. USA* **98**, 14487–14492 (2001).
- Ganguly, A., Jiang, J. & Ip, Y. T. *Drosophila* WntD is a target and an inhibitor of the Dorsal/Twist/Snail network in the gastrulating embryo. *Development* **132**, 3419–3429 (2005).
- Gillespie, S. K. & Wasserman, S. A. Dorsal, a *Drosophila* Rel-like protein, is phosphorylated upon activation of the transmembrane protein Toll. *Mol. Cell. Biol.* **14**, 3559–3568 (1994).
- Drier, E. A., Govind, S. & Steward, R. Cactus-independent regulation of Dorsal nuclear import by the ventral signal. *Curr. Biol.* **10**, 23–26 (2000).
- Logan, C. Y. & Nusse, R. The Wnt signalling pathway in development and disease. *Annu. Rev. Cell Dev. Biol.* **20**, 781–810 (2004).
- Hamada, F. et al. Negative regulation of Wingless signalling by D-axin, a *Drosophila* homolog of axin. *Science* **283**, 1739–1742 (1999).
- Rusch, J. & Levine, M. Regulation of the dorsal morphogen by the Toll and torso signalling pathways: a receptor tyrosine kinase selectively masks transcriptional repression. *Genes Dev.* **8**, 1247–1257 (1994).
- Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J. M. & Hoffmann, J. A. The dorsoventral regulatory gene cassette *spatzle*/Toll/cactus controls the potent antifungal response in *Drosophila* adults. *Cell* **86**, 973–983 (1996).
- Meng, X., Khanuja, B. S. & Ip, Y. T. Toll receptor-mediated *Drosophila* immune response requires Dif, an NF- κ B factor. *Genes Dev.* **13**, 792–797 (1999).
- Choe, K. M., Werner, T., Stoven, S., Hultmark, D. & Anderson, K. V. Requirement for a peptidoglycan recognition protein (PGRP) in Relish activation and antibacterial immune responses in *Drosophila*. *Science* **296**, 359–362 (2002).
- Hedengren, M. et al. Relish, a central factor in the control of humoral but not cellular immunity in *Drosophila*. *Mol. Cell* **4**, 827–837 (1999).
- De Gregorio, E., Spellman, P. T., Tzou, P., Rubin, G. M. & Lemaitre, B. The Toll and Imd pathways are the major regulators of the immune response in *Drosophila*. *EMBO J.* **21**, 2568–2579 (2002).
- Green, C. et al. The necrotic gene in *Drosophila* corresponds to one of a cluster of three serpin transcripts mapping at 43A1.2. *Genetics* **156**, 1117–1127 (2000).
- Lemaitre, B. et al. Functional analysis and regulation of nuclear import of dorsal during the immune response in *Drosophila*. *EMBO J.* **14**, 536–545 (1995).
- Bettencourt, R., Asha, H., Dearolf, C. & Ip, Y. T. Hemolymph-dependent and -independent responses in *Drosophila* immune tissue. *J. Cell. Biochem.* **92**, 849–863 (2004).
- Lemaitre, B. et al. A recessive mutation, immune deficiency (*imd*), defines two distinct control pathways in the *Drosophila* host defense. *Proc. Natl Acad. Sci. USA* **92**, 9465–9469 (1995).
- Gross, I., Georgel, P., Kappler, C., Reichhart, J. M. & Hoffmann, J. A. *Drosophila* immunity: a comparative analysis of the Rel proteins dorsal and Dif in the induction of the genes encoding dipterin and cecropin. *Nucleic Acids Res.* **24**, 1238–1245 (1996).
- Mansfield, B. E., Dionne, M. S., Schneider, D. S. & Freitag, N. E. Exploration of host-pathogen interactions using *Listeria monocytogenes* and *Drosophila melanogaster*. *Cell. Microbiol.* **5**, 901–911 (2003).
- Brandt, S. M. et al. Secreted bacterial effectors and host-produced Eiger/TNF drive death in a *Salmonella*-infected fruit fly. *PLoS Biol.* **2**, e418 (2004).
- Zambon, R. A., Nandakumar, M., Vakharia, V. N. & Wu, L. P. The Toll pathway is important for an antiviral response in *Drosophila*. *Proc. Natl Acad. Sci. USA* **102**, 7257–7262 (2005).
- Zettervall, C. J. et al. A directed screen for genes involved in *Drosophila* blood cell activation. *Proc. Natl Acad. Sci. USA* **101**, 14192–14197 (2004).
- Rorth, P. Gal4 in the *Drosophila* female germline. *Mech. Dev.* **78**, 113–118 (1998).
- Dionne, M. S., Ghori, N. & Schneider, D. S. *Drosophila melanogaster* is a genetically tractable model host for *Mycobacterium marinum*. *Infect. Immun.* **71**, 3540–3550 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Brown for preparing WntD protein for antibody production; C.-H. Wu and K. Xu for preliminary experiments with *wntD*; M. Fish for p-element transformations; J. Sekelsky and K. Golic for fly strains and vectors used in gene targeting; S. Roth, G. Zimmermann and the Developmental Studies Hybridoma Bank for antibodies; A. O'Reilly and the Bloomington Stock Center for fly stocks; and C. Logan and A. Mikels for reading of the manuscript. M.D.G. was supported by a Howard Hughes Medical Institute predoctoral fellowship and a Stanford Graduate Fellowship. This work was supported by grants from the NIH to R.N. and D.S.S. and by the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.N. (rnusse@stanford.edu).

LETTERS

Interference with AI-2-mediated bacterial cell-cell communication

Karina B. Xavier¹ & Bonnie L. Bassler¹

Bacteria communicate by means of chemical signal molecules called autoinducers. This process, called quorum sensing, allows bacteria to count the members in the community and to alter gene expression synchronously across the population. Quorum-sensing-controlled processes are often crucial for successful bacterial–host relationships—both symbiotic and pathogenic. Most quorum-sensing autoinducers promote intraspecies communication, but one autoinducer, called AI-2, is produced and detected by a wide variety of bacteria and is proposed to allow interspecies communication^{1,2}. Here we show that some species of bacteria can manipulate AI-2 signalling and interfere with other species' ability to assess and respond correctly to changes in cell population density. AI-2 signalling, and the interference with it, could have important ramifications for eukaryotes in the maintenance of normal microflora and in protection from pathogenic bacteria.

The bacterial signal molecule autoinducer-2 (AI-2) is a product of the LuxS enzyme, which is widely conserved throughout the bacterial kingdom. LuxS enzymes synthesize 4,5-dihydroxy-2,3-pentanedione (DPD), which undergoes spontaneous rearrangements³ to form DPD derivatives that interconvert and exist in equilibrium. Different bacteria recognize distinct DPD derivatives, and this family of molecules is generically referred to as AI-2 (ref. 4). The interconverting nature of these molecules presumably allows bacteria to respond to their own AI-2 and also to AI-2 produced by other bacterial species, suggesting that AI-2 represents a universal language allowing interspecies bacterial communication.

We previously characterized the quorum-sensing signal production and detection mechanisms in *Vibrio harveyi*, *V. cholerae*, *Escherichia coli* and *Salmonella typhimurium* (refs 5–8). In *V. harveyi*, two autoinducers, AI-1 and AI-2, are detected by LuxN and LuxPQ, respectively, and control the expression of genes including those for bioluminescence and type III secretion (TTS) of virulence factors (Fig. 1a; refs 5, 9). A related quorum-sensing network exists in *V. cholerae* and controls the expression of virulence genes including *hapA*, encoding the haemagglutinin (H/A) protease^{6,10}.

Some bacteria both produce and consume AI-2. For example, *E. coli* and *S. typhimurium* release AI-2 in exponential phase, and import AI-2 at the transition into stationary phase. This occurs because one target that is activated by AI-2 is the Lsr (for LuxS-regulated) transporter that imports AI-2 (Fig. 1b; refs 7, 8). In the absence of AI-2, LsrR represses the *lsr* operon. Following AI-2 release, low-level internalization occurs and the intracellular AI-2 is then phosphorylated by LsrK (refs 8, 11). Phosphorylated AI-2 antagonizes LsrR, which leads to de-repression of *lsr* expression, assembly of the Lsr transporter and rapid AI-2 internalization. LsrR[−] strains are avid AI-2 consumers because the *lsr* operon is de-repressed. By contrast, LsrK[−] strains never consume significant amounts of AI-2 due to their inability to de-repress *lsr* transcription (Fig. 1b, Supplementary Figs 1S and 2S; refs 8, 11).

In mixed-species consortia, production and consumption of AI-2 by enteric bacteria should have reciprocal effects on gene regulation in other bacterial species that communicate using AI-2. When enteric bacteria supply AI-2, a nearby species could use the information to count the enteric cells in the mixed-species community or prematurely activate the expression of quorum-sensing-regulated genes. However, when enteric bacteria remove AI-2, a neighbouring species could underestimate cell number and fail to initiate or incorrectly terminate quorum sensing. Thus, consuming AI-2 could allow enteric cells to interfere with AI-2-mediated communication in other bacteria.

When *E. coli* is grown in pure culture, AI-2 is required to activate *lsr* expression, as shown by comparing wild-type cells (first bar in Fig. 2a) with LuxS[−] cells (second bar in Fig. 2a). Induction of *lsr* expression also occurs when LuxS[−] *E. coli* is mixed with wild-type *V. harveyi* (third bar in Fig. 2a), and depends on the AI-2 supplied by *V. harveyi* because no *lsr* activation occurs when LuxS[−] *E. coli* is combined with LuxS[−] *V. harveyi* (fourth bar in Fig. 2a). Thus, *E. coli* detects and responds to its own AI-2 and also to AI-2 produced by *V. harveyi*.

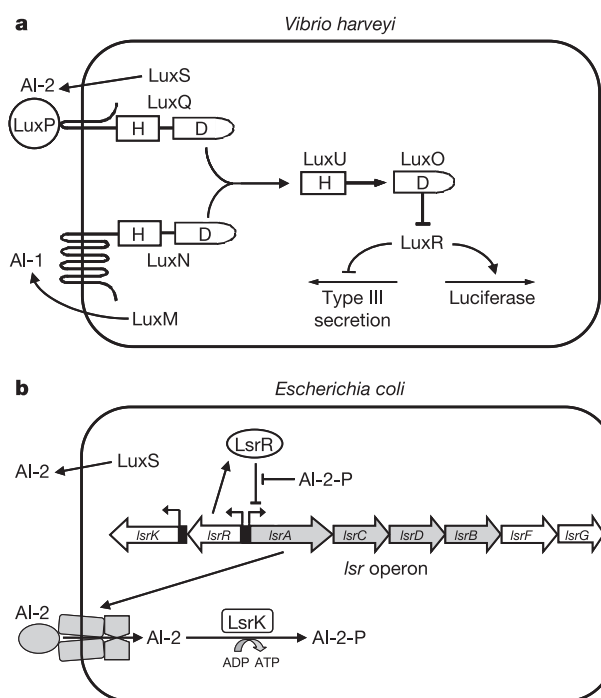


Figure 1 | AI-2 signalling systems. a, *V. harveyi* quorum sensing. The autoinducers AI-1 and AI-2 are detected by LuxN and LuxPQ, respectively. Information is transduced by phosphorylation. **b, The *E. coli* Lsr transporter imports AI-2.** See the text for details. AI-2-P, phosphorylated AI-2.

¹Department of Molecular Biology, Howard Hughes Medical Institute, Princeton University, Princeton, New Jersey 08544-1014, USA.

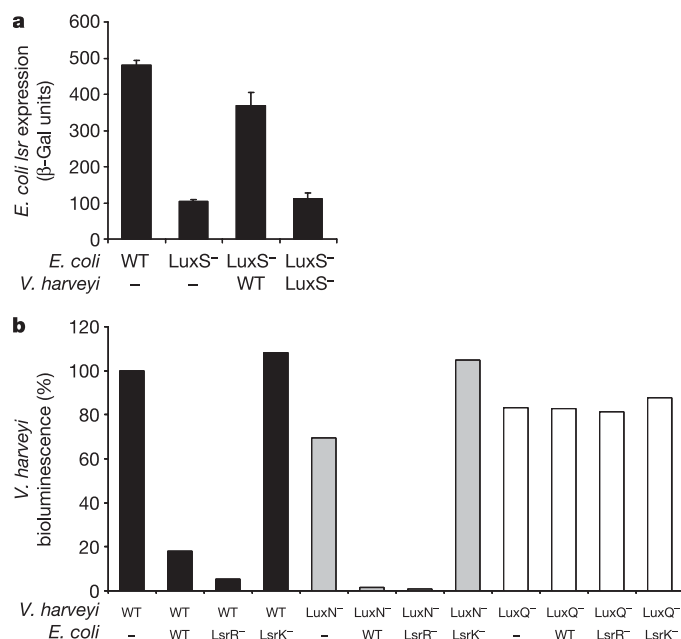


Figure 2 | *E. coli* and *V. harveyi* communicate using AI-2. **a**, *V. harveyi* AI-2 induces the *E. coli* *lsr* operon. β-Galactosidase activities of an *E. coli* *lsr-lacZ* fusion in pure *E. coli* cultures and co-cultures with *V. harveyi*. Wild type (WT) cells are AI-2⁺; LuxS⁻ cells are AI-2⁻. β-Gal units are defined as (A₄₂₀ min⁻¹ × 10⁹ ml⁻¹)/(*E. coli* c.f.u. ml⁻¹). Error bars represent standard deviations. **b**, *E. coli* consumes AI-2 and inhibits *V. harveyi* bioluminescence expression. Light production in *V. harveyi* wild-type (black bars), LuxN⁻ (grey bars) and LuxQ⁻ (white bars) cells in co-culture with *E. coli* wild-type, LsrR⁻ and LsrK⁻ cells. *V. harveyi* bioluminescence is expressed as a percentage of wild-type bioluminescence (140,000 relative light units, r.l.u., in units of (c.p.m. ml⁻¹ × 10³)/(*V. harveyi* c.f.u. ml⁻¹).

We also examined the effects of *E. coli* AI-2 production and consumption on the behaviour of *V. harveyi* using bioluminescence as an indicator. When incubated with wild-type *E. coli*, wild-type *V. harveyi* at high cell density produces only 18% of the bioluminescence (light) it produces in pure culture (first and second black bars in Fig. 2b). Thus, *E. coli* consumes both its own AI-2 and the AI-2 of *V. harveyi*, causing a reduction in light output from *V. harveyi*. Co-culture of wild-type *V. harveyi* with LsrR⁻ *E. coli* (a constitutive AI-2 importer) causes a greater reduction in *V. harveyi* bioluminescence

(third black bar in Fig. 2b). Diminution of light is due exclusively to Lsr-mediated transport of AI-2 into *E. coli* because no reduction in light production occurs when *V. harveyi* is co-cultured with the LsrK⁻ *E. coli* mutant that does not internalize AI-2 (fourth black bar in Fig. 2b).

In addition to AI-2, *V. harveyi* produces an autoinducer called AI-1 and responds to it through the LuxN sensor (Fig. 1a; ref. 5). *V. harveyi* LuxN⁻ strains produce less light than wild-type cells because LuxN⁻ strains induce bioluminescence only in response to AI-2, whereas wild-type cells respond to both AI-1 and AI-2 (first grey bar in Fig. 2b). Growth of LuxN⁻ *V. harveyi* with wild-type and LsrR⁻ *E. coli* reduces light output to 2% and <1%, respectively (second and third grey bars in Fig. 2b). *E. coli* has a stronger effect on LuxN⁻ *V. harveyi* than on wild-type *V. harveyi* because wild-type *V. harveyi* retains its response to AI-1, the levels of which are not altered by *E. coli* consumption of AI-2. When LuxN⁻ *V. harveyi* is co-cultured with the LsrK⁻ *E. coli* mutant, increased light production occurs in response to the additional AI-2 provided by the internalization-defective *E. coli* cells (fourth grey bar in Fig. 2b). Finally, these effects of AI-2 on *V. harveyi* require the LuxQ sensor (Fig. 1a) because manipulating extracellular AI-2 levels by co-culture with *E. coli* does not alter light production by the LuxQ⁻ *V. harveyi* (white bars in Fig. 2b).

Dilution of *V. harveyi* in pure culture causes autoinducer levels to fall below those required for detection, and bioluminescence expression terminates at low cell densities. During subsequent growth, autoinducers are again released and, on achieving a threshold concentration, they are detected and the cells respond by inducing a rapid increase in light production. Our premise is that in mixed cultures, early production and later consumption of AI-2 by *E. coli* should inversely affect the quorum-sensing response of *V. harveyi*. To test this idea, we measured *V. harveyi* bioluminescence during growth in co-culture with different strains of *E. coli*. Both species were diluted to a low cell density and combined under conditions allowing each to grow exponentially (Fig. 3a). *V. harveyi* grown with the non-AI-2-producing, non-AI-2-importing *E. coli* LuxS⁻, LsrK⁻ strain shows the characteristic initial decline in bioluminescence followed by the rapid increase to the maximal, pre-dilution level of light production (crosses in Fig. 3b). In the presence of all AI-2-producing *E. coli* strains tested, the initial decrease in bioluminescence is greatly reduced because *V. harveyi* rapidly initiates quorum sensing in response to the AI-2 supplied by *E. coli* (filled squares, plus signs and filled triangles in Fig. 3b). At later time points, corresponding to post-*lsr*-induction, all *E. coli* strains

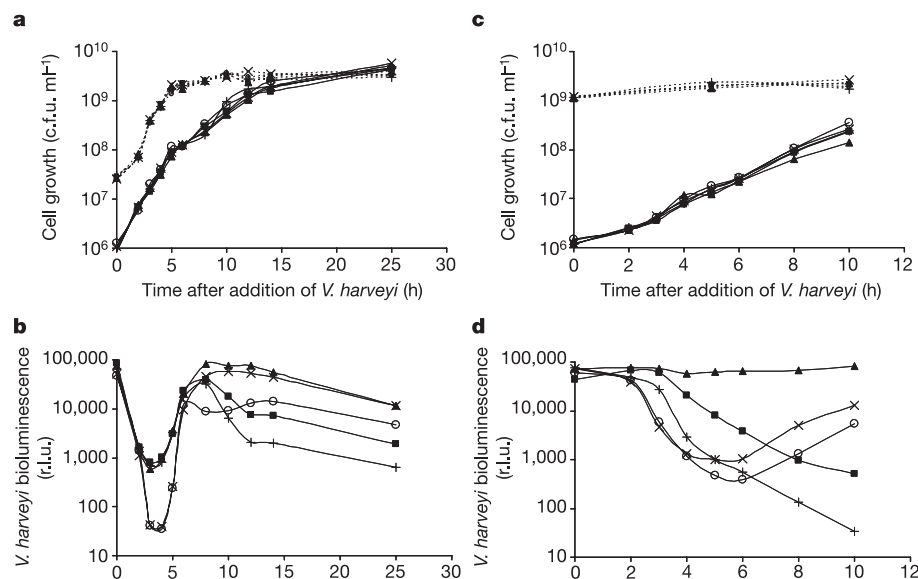


Figure 3 | The *E. coli* Lsr system interferes with *V. harveyi* quorum sensing. **a**, Growth curves for strains shown in panel b. Solid and dotted lines denote *V. harveyi* and *E. coli* c.f.u., respectively. **b**, Light production for LuxN⁻ *V. harveyi* in co-culture with the *E. coli* strains wild type (filled squares), LsrR⁻ (plus signs), LsrK⁻ (filled triangles), LuxS⁻ (open circles) and LsrK⁻, LuxS⁻ (crosses). **c**, Growth curves for strains shown in panel d. Solid and dotted lines denote *V. harveyi* and *E. coli* c.f.u., respectively. **d**, LuxN⁻ *V. harveyi* was diluted into *E. coli* cultures entering stationary phase. *E. coli* strains are identical to those shown in panel b.

that have LsrK, and are therefore capable of transporting AI-2, consume the AI-2 (filled squares, plus signs and open circles in Fig. 3b). This has the consequence of decreasing *V. harveyi* light output by almost 100-fold compared with *V. harveyi* grown in the presence of LsrK⁻ strains that are unable to internalize AI-2 (filled triangles and crosses in Fig. 3b). Mixing *V. harveyi* with the *E. coli* LsrR⁻ strain results in the most pronounced reduction in light production, presumably owing to the constitutive removal of AI-2 (plus signs in Fig. 3b).

A more marked effect on *V. harveyi* quorum sensing occurs when exponential-phase *V. harveyi* encounters the *E. coli* strains in stationary phase (Fig. 3c). In this case, at the time of mixing wild-type *E. coli* is induced for *lsr* expression, and the cells are actively internalizing AI-2. An almost immediate and continuous decrease in *V. harveyi* bioluminescence occurs (filled squares in Fig. 3d), which is even more obvious in the mixture containing the LsrR⁻ *E. coli* strain (plus signs in Fig. 3d). In contrast, co-culture with stationary-phase LsrK⁻ *E. coli* results in a steady maintenance of maximal light production in *V. harveyi* (filled triangles in Fig. 3d), owing to the presence of AI-2 produced by the transport-deficient *E. coli* mutant. As expected, in the presence of the LsrK⁻,LuxS⁻ *E. coli* strain, *V. harveyi* shows density-dependent bioluminescence (crosses in Fig. 3d). Mixing *V. harveyi* with LuxS⁻ *E. coli* cells initially has no effect on *V. harveyi* quorum sensing because the LuxS⁻ *E. coli* strain has not been exposed to AI-2 before the addition of *V. harveyi*, so the Lsr transporter is repressed (open circles in Fig. 3d). During co-incubation of these bacteria, the *V. harveyi*-produced AI-2 induces LuxS⁻ *E. coli* cells to express the Lsr transporter. *E. coli* cells then consume AI-2, causing a reduction in *V. harveyi* light production. Therefore, at later time points, light output from the *V. harveyi* strain

mixed with the *E. coli* LuxS⁻ strain falls below that of *V. harveyi* co-incubated with the *E. coli* LsrK⁻,LuxS⁻ strain.

The *V. harveyi* quorum-sensing regulon contains many genes, and, presumably, all are susceptible to the effects of AI-2 production and consumption within consortia. To examine this possibility, the expression of the TTS gene *vopN* was measured in co-cultures of different *E. coli* and *V. harveyi* strains. TTS genes are repressed by autoinducers at high cell density in *V. harveyi* (Fig. 1a; ref. 9). Production of AI-2 by an *E. coli* strain incapable of importing AI-2 does not affect *vopN* expression, showing that AI-2 supplied by *V. harveyi* is sufficient for maximal repression of *vopN* (black bars in Fig. 4a; compare LsrK⁻ cells to LsrK⁻,LuxS⁻ cells). Growth with *E. coli* strains that consume AI-2 results in de-repression of *vopN* expression (black bars in Fig. 4a; compare wild-type cells to LsrR⁻ cells). Full de-repression occurs only in combinations where neither *V. harveyi* nor *E. coli* produces AI-2 (hatched bars in Fig. 4a), demonstrating that in the previous mixtures (depicted by the black bars in Fig. 4a) *E. coli* has not internalized all of the AI-2. As in Fig. 2a, AI-2 production and consumption does not affect *vopN* expression if *V. harveyi* lacks LuxQ (white bars in Fig. 4a).

We recognize that *V. harveyi*, which routinely lives in mixed-species marine communities, is unlikely to coexist in the same environment as *E. coli*. In contrast, *V. cholerae* must coexist with *E. coli* during pathogenic associations with humans. To determine whether *E. coli* AI-2 consumption could interfere with *V. cholerae* signalling, we measured the expression of a gene that is activated during quorum sensing: *hapA*, encoding the H/A protease. These results mimic those of TTS gene expression in *V. harveyi* (Fig. 4a) except that the pattern of regulation is reversed (Fig. 4b). Induction of gene expression occurs in mixtures containing *E. coli* strains incapable of AI-2 transport (black bars in Fig. 4b; compare LsrK⁻ cells to LsrK⁻,LuxS⁻ cells) and induction of gene expression does not occur in mixtures containing *E. coli* strains that can transport AI-2 (black bars in Fig. 4b; compare wild-type cells to LsrR⁻ cells).

AI-2 consumption has a more modest effect on *vopN* and *hapA* regulation than on the regulation of bioluminescence. Many environmental factors are known to control H/A protease production and TTS (refs 10, 12). AI-2 interference probably has more subtle effects on H/A protease levels and TTS than on bioluminescence because additional regulators exert control over H/A protease and TTS output, and regulation by these factors remains unchanged in our experiments.

Our results show that AI-2 can mediate two-way communication between bacterial species, because when *E. coli* and *V. harveyi* are grown in co-culture, AI-2 production by either species can regulate light production in *V. harveyi* and trigger *lsr* induction in *E. coli*. Thus, AI-2 production allows each species to include the other in the 'census'. Because *E. coli* and *V. harveyi* respond to AI-2 signals with different structures^{2,4}, cross-communication implies that the signal released by one species must convert into that used by the other species. Presumably, these transformations occur in the medium between the sender and receiver cells. This is of particular interest given that *Vibrio* species detect an AI-2 molecule containing boron, whereas there is no boron in the AI-2 signal recognized by enteric bacteria^{2,4}. Our results demonstrate that these marked chemical interconversions are occurring on a timescale that promotes major effects on gene expression.

In model *V. harveyi*-*E. coli* mixtures, induction of *lsr* genes in *E. coli* results in the assembly of the AI-2 transporter and subsequent consumption of AI-2. This has the consequence of inhibiting light production from *V. harveyi*, demonstrating that *E. coli* interferes with AI-2-mediated communication by causing *V. harveyi* to terminate prematurely its quorum-sensing behaviours. The impact of the *E. coli* AI-2-Lsr system is not restricted to interfering with the activation of bioluminescence because repression of TTS is also affected. These findings imply that interference with AI-2 signalling influences the expression of entire quorum-sensing regulons.

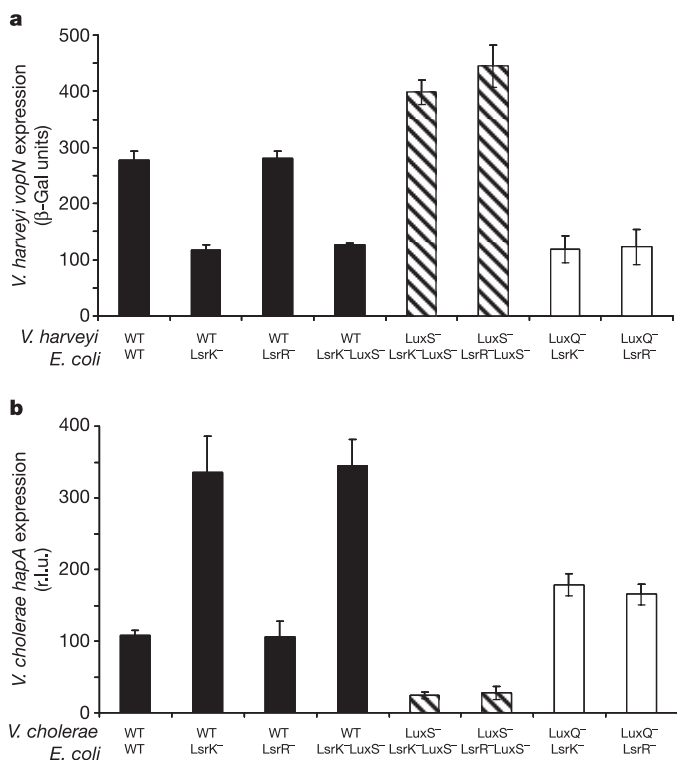


Figure 4 | *V. harveyi* TTS and *V. cholerae* H/A protease production are regulated by *E. coli* Lsr-mediated AI-2 transport. **a**, *V. harveyi* TTS expression was measured using a *vopN-lacZ* fusion (18 h co-culture). **b**, *V. cholerae* H/A protease expression was measured with a *hapA-lux* fusion (16 h co-culture). Activities were measured in the *Vibrio* strains wild type (black bars), LuxS⁻ (hatched bars) and LuxQ⁻ (white bars). Each *Vibrio* strain was grown in co-culture with the *E. coli* strains wild type, LsrK⁻, LsrR⁻, and the two double mutants LsrK⁻,LuxS⁻, and LsrR⁻,LuxS⁻. Error bars represent standard deviations.

Lsr-mediated interference with AI-2 signalling also occurs in co-cultures of *E. coli* and *V. cholerae*, two bacteria that certainly co-colonize the human intestine during *V. cholerae* infection. We do not expect these interactions to be restricted to enteric bacteria; in fact, the disappearance of AI-2 in stationary phase has been reported for diverse bacterial species, indicating that either Lsr transporters exist in these species, or that additional mechanisms exist to eliminate AI-2 (ref. 13). Any species of bacteria that relies on AI-2-mediated communication and inhabits niches containing another bacterial species that produces and/or consumes AI-2 could be similarly affected. Whether quorum sensing is enhanced or inhibited will depend on the growth status of the different species when they encounter one another. Complex pro- and anti-AI-2-mediated interactions could be taking place in natural niches. It is also possible that eukaryotes have capitalized on this by evolving specific associations with bacteria that use or manipulate AI-2 signalling. These interactions could have important consequences for humans in the maintenance of normal gut microflora, in addition to the prevention of bacterial diseases.

METHODS

Bacterial strains and growth conditions. Bioluminescence was measured in *V. harveyi* strain BB120 (wild type; ref. 1), MM30 (*luxS*[−]; ref. 14), BB170 (*luxN*[−]; ref. 15) and BB960 (*luxQ*[−]; ref. 5). *E. coli* strains were derivatives of MG1655 (ref. 16). The *E. coli* strains used in Figs 2 and 3, and Supplementary Figs 1S and 2S, contain a *lacZYA* deletion and an *lsr-lacZ* fusion integrated at the *att* site. These strains⁸ are KX1123 (wild type), KX1218 (*luxS*[−]), KX1186 (*lsrK*[−]), KX1322 (*lsrR*[−]) and KX1372 (*lsrK*[−], *luxS*[−]). To measure *vopN* expression, *V. harveyi* strains JMH385 (wild type), KX1530 (*luxS*[−]) and JMH669 (*luxQ*[−]), containing a chromosomal *vopN::mini-MulacZ* insertion, were used⁹. The final two strains were obtained by the introduction of *luxS::Tn5* (ref. 14) or *luxQ::Tn5* (ref. 5) onto the chromosome of JMH385, respectively. *E. coli* strains used in TTS experiments are KX1102 (wild type), KX1479 (*lsrK*[−]), KX1477 (*lsrR*[−]) and KX1526 (*lsrK*[−], *luxS*[−]), and each was constructed by the method reported in ref. 8. *V. cholerae* strains are derivatives of El Tor C6706str2, a streptomycin-resistant isolate of C6706 (ref. 17), and all have a *hap-luxCDABE* transcriptional fusion cloned into a plasmid containing a chloramphenicol resistance (Cm^r) marker. The *hap-luxCDABE* transcriptional fusion was constructed as reported¹⁸, except that a PCR-amplified fragment containing the promoter region of *V. cholerae* *hapA* was cloned into unique *SpeI* to *BamHI* sites. The *V. cholerae* strains used are: BH1220 (wild type), BH1312 (*luxS*[−]) and BH1253 (*luxQ*[−]). Details of their construction will be reported elsewhere (B. K. Hammer and B.L.B., manuscript in preparation). *E. coli* strains used for co-culture with *V. cholerae* are identical to those described for Fig. 3 except for KX1583 (wild type), which contains a Cm^r marker downstream of the *luxS* gene. Addition of chloramphenicol was required in *E. coli*–*V. cholerae* co-cultures to maintain the plasmid harbouring the *hapA-luxCDABE* fusion in *V. cholerae*. Co-cultures were grown in Luria–Marine (LM) medium (ref. 5) in experiments with *V. harveyi*, or in Luria–Bertani (LB) medium supplemented with chloramphenicol (10 mg l^{−1}) in experiments with *V. cholerae*. All cultures were incubated at 30 °C with aeration.

β-Galactosidase assays. Overnight cultures of *E. coli* and *V. harveyi* strains were diluted 1:1,000 into LM and grown for 8 h, after which time *lsr-lacZ* expression was measured. When *vopN-lacZ* expression was measured, *V. harveyi* strains were diluted 1:5,000 into LM and 1:100 dilutions of overnight cultures of *E. coli* were added. Cultures were grown for 18 h because expression of TTS genes is maximal in stationary phase. Cells from 1 ml of culture were resuspended in 1 ml of Z buffer for β-galactosidase assays¹⁹. The values shown represent the average of triplicate experiments.

Bioluminescence assays. Overnight cultures of *V. harveyi* strains were diluted 1:5,000 into LM. In co-incubation experiments, 1:100 dilutions of overnight cultures of *E. coli* were added. In the experiments shown in Fig. 2b, bioluminescence was measured after 12 h of incubation. In the experiments shown in Fig. 3c and d, the *V. harveyi* strains were added to *E. coli* cultures that had been pre-grown for 4 h in LM. At the reported times, aliquots were plated to determine the number of colony-forming units (c.f.u.), and bioluminescence was measured using a liquid scintillation counter (Wallac model 1409, PerkinElmer).

Analysis of *hapA-luxCDABE* expression. Overnight cultures of *E. coli* and *V. cholerae* were diluted 1:1,000 into LB supplemented with chloramphenicol and grown for 16 h. Aliquots were plated to determine c.f.u., and luciferase activity from the *hapA-luxCDABE* fusion was measured in a liquid scintillation counter. Values shown represent the average of triplicate experiments.

Measurements of cell number in co-cultures. In all co-culture experiments, the mixtures of cells were plated onto two different types of media. The media conditions were chosen so that they either permitted growth of only one of the two species in the mixture, or the colonies of each species could be distinguished visually by morphology. This strategy allowed us to determine the cell numbers of both species in each co-culture experiment. *V. harveyi* c.f.u. were assessed following overnight incubation at 30 °C on LM agar supplemented with ampicillin (100 mg l^{−1}). *V. cholerae* c.f.u. were counted after incubation at 37 °C on LB agar containing 50 mg l^{−1} polymyxin B. *E. coli* c.f.u. were counted following incubation at 37 °C on LB. *V. harveyi* does not grow at 37 °C and *V. cholerae* colonies are translucent under this condition, whereas *E. coli* colonies are opaque.

Received 29 April; accepted 27 June 2005.

- Bassler, B. L., Greenberg, E. P. & Stevens, A. M. Cross-species induction of luminescence in the quorum-sensing bacterium *Vibrio harveyi*. *J. Bacteriol.* **179**, 4043–4045 (1997).
- Chen, X. *et al.* Structural identification of a bacterial quorum-sensing signal containing boron. *Nature* **415**, 545–549 (2002).
- Schauder, S., Shokat, K., Surette, M. G. & Bassler, B. L. The LuxS family of bacterial autoinducers: biosynthesis of a novel quorum-sensing signal molecule. *Mol. Microbiol.* **41**, 463–476 (2001).
- Miller, S. T. *et al.* *Salmonella typhimurium* recognizes a chemically distinct form of the bacterial quorum-sensing signal AI-2. *Mol. Cell* **15**, 677–687 (2004).
- Bassler, B. L., Wright, M. & Silverman, M. R. Multiple signalling systems controlling expression of luminescence in *Vibrio harveyi*: sequence and function of genes encoding a second sensory pathway. *Mol. Microbiol.* **13**, 273–286 (1994).
- Miller, M. B., Skorupski, K., Lenz, D. H., Taylor, R. K. & Bassler, B. L. Parallel quorum sensing systems converge to regulate virulence in *Vibrio cholerae*. *Cell* **110**, 303–314 (2002).
- Taga, M. E., Semmelhack, J. L. & Bassler, B. L. The LuxS-dependent autoinducer AI-2 controls the expression of an ABC transporter that functions in AI-2 uptake in *Salmonella typhimurium*. *Mol. Microbiol.* **42**, 777–793 (2001).
- Xavier, K. B. & Bassler, B. L. Regulation of uptake and processing of the quorum-sensing autoinducer AI-2 in *Escherichia coli*. *J. Bacteriol.* **187**, 238–248 (2005).
- Henke, J. M. & Bassler, B. L. Quorum sensing regulates type III secretion in *Vibrio harveyi* and *Vibrio parahaemolyticus*. *J. Bacteriol.* **186**, 3794–3805 (2004).
- Silva, A. J. & Benitez, J. A. Transcriptional regulation of *Vibrio cholerae* hemagglutinin/protease by the cyclic AMP receptor protein and RpoS. *J. Bacteriol.* **186**, 6374–6382 (2004).
- Taga, M. E., Miller, S. T. & Bassler, B. L. Lsr-mediated transport and processing of AI-2 in *Salmonella typhimurium*. *Mol. Microbiol.* **50**, 1411–1427 (2003).
- Hacker, J. & Kaper, J. B. Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* **54**, 641–679 (2000).
- Xavier, K. B. & Bassler, B. L. LuxS quorum sensing: more than just a numbers game. *Curr. Opin. Microbiol.* **6**, 191–197 (2003).
- Surette, M. G., Miller, M. B. & Bassler, B. L. Quorum sensing in *Escherichia coli*, *Salmonella typhimurium*, and *Vibrio harveyi*: a new family of genes responsible for autoinducer production. *Proc. Natl Acad. Sci. USA* **96**, 1639–1644 (1999).
- Bassler, B. L., Wright, M., Showalter, R. E. & Silverman, M. R. Intercellular signalling in *Vibrio harveyi*: sequence and function of genes regulating expression of luminescence. *Mol. Microbiol.* **9**, 773–786 (1993).
- Daniels, D. L., Plunkett, G. III, Burland, V. & Blattner, F. R. Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. *Science* **257**, 771–778 (1992).
- Thelin, K. H. & Taylor, R. K. Toxin-coregulated pilus, but not mannose-sensitive hemagglutinin, is required for colonization by *Vibrio cholerae* O1 El Tor biotype and O139 strains. *Infect. Immun.* **64**, 2853–2856 (1996).
- Lenz, D. H. *et al.* The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholerae*. *Cell* **118**, 69–82 (2004).
- Slauch, J. M. & Silhavy, T. J. *cis*-acting *ompF* mutations that result in OmpR-dependent constitutive expression. *J. Bacteriol.* **173**, 4039–4048 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was supported by HHMI and NIH and NSF grants (B.L.B.) and a Praxis XXI postdoctoral fellowship, Portugal (K.B.X.). We thank the members of the Bassler laboratory, as well as N. Wingreen and F. Hughson for insightful discussions. We are grateful to B. Hammer for *V. cholerae* strains and to J. Henke for *V. harveyi* TTS mutants.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.L.B. (bbassler@molbio.princeton.edu).

LETTERS

Phosphatidylserine-dependent engulfment by macrophages of nuclei from erythroid precursor cells

Hideyuki Yoshida¹, Kohki Kawane^{1,3,4}, Masato Koike², Yoshimi Mori⁵, Yasuo Uchiyama² & Shigekazu Nagata^{1,3,4}

Definitive erythropoiesis usually occurs in the bone marrow or fetal liver, where erythroblasts are associated with a central macrophage in anatomical units called 'blood islands'^{1,2}. Late in erythropoiesis, nuclei are expelled from the erythroid precursor cells and engulfed by the macrophages in the blood island^{2,3}. Here we show that the nuclei are engulfed by macrophages only after they are disconnected from reticulocytes, and that phosphatidylserine, which is often used as an 'eat me' signal for apoptotic cells, is also used for the engulfment of nuclei expelled from erythroblasts. We investigated the mechanism behind the enucleation and engulfment processes by isolating late-stage erythroblasts from the spleens of phlebotomized mice. When these erythroblasts were cultured, the nuclei protruded spontaneously from the erythroblasts. A weak physical force could disconnect the nuclei from the reticulocytes. The released nuclei contained an undetectable level of ATP, and quickly exposed phosphatidylserine on their surface. Fetal liver macrophages efficiently engulfed the nuclei; masking the phosphatidylserine on the nuclei with the dominant-negative form of milk-fat-globule EGF8 (MFG-E8) prevented this engulfment.

Phlebotomy promotes the development of erythroid cells in mouse spleen^{4,5}. We used this system to obtain a large number of erythroblasts at a synchronized developmental stage. When prepared from spleens 4 days after phlebotomy, about 60% of the purified erythroblasts were CD71⁺Ter119⁺ (Fig. 1a). Few CD117(c-Kit)⁺ cells were present, indicating that the cells were erythroblasts at the late stage of erythropoiesis (Fig. 1b)⁶. When the erythroblasts were cultured at 37 °C, some cells started blebbing and the nucleus moved to one side of the cell (Supplementary Movie 1). After 5 h, about 12% of the erythroblasts possessed an eccentric nucleus that could be distinguished from the irregularly shaped reticulocytes (Fig. 1c). Observation by electron microscopy indicated that the nucleus was surrounded by an intact plasma membrane and was connected to the reticulocytes by means of a thin pseudopodium-like membrane extension (Fig. 1d). After collection and resuspension of the enucleating erythroblasts, most nuclei became disconnected from the reticulocytes, indicating that the connection between the nucleus and reticulocyte could be disrupted by weak physical stress. In fact, many nuclei could be disconnected from reticulocytes by gently pipetting the erythroblasts in and out five times.

We then monitored the enucleation process with two different DNA-staining dyes, SYTO16 and SYTOX blue, which are cell-permeable and cell-impermeable fluorescent dyes, respectively. Erythroblasts were represented by a population (P6) of high forward scatter (FSC) with SYTO16^{high} SYTOX⁻ (Fig. 2a). The cells with low FSC (FSC^{low}) consist of two populations, the expelled nuclei, which were SYTO16^{high} SYTOX⁻ (P3), and SYTO16^{low} SYTOX⁻ reticulocytes (P4). All SYTOX⁺ cells were dead. When the erythroblasts were

incubated at 37 °C, their number did not change significantly (Fig. 2b). There were no free nuclei before incubation, but their number increased gradually during incubation. This increase paralleled an increase in reticulocyte numbers. Because the extruded

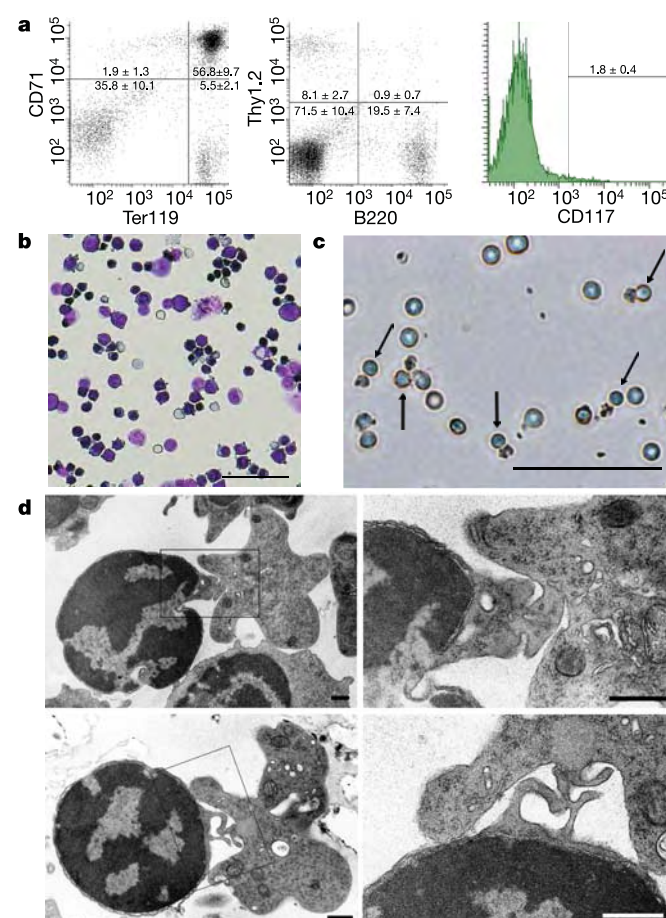


Figure 1 | Enucleation of erythroblasts. **a**, Erythroblasts from phlebotomized mice were stained for Ter119 and CD71, or B220 and Thy 1.2, and analysed by flow cytometry. Numbers indicate the percentage of stained cells in each quadrant. Experiments were performed three times, and the numerical results shown are means ± s.d. The staining profile for CD117 (c-Kit) is also shown. **b**, Erythroblasts stained with Wright-Giemsa. Scale bar, 50 μm. **c**, Erythroblasts incubated for 5 h. Nuclei protruding from erythroblasts are indicated by arrows. Scale bar, 50 μm. **d**, Electron microscopy of erythroblasts incubated for 3 h. Boxed areas in the left panels are enlarged in the right panels. Scale bar, 1 μm.

¹Department of Genetics and ²Department of Cell Biology and Neuroscience, Graduate School of Medicine, Osaka University, Osaka 565-0871, Japan. ³Laboratory of Genetics, Integrated Biology Laboratories, Graduate School of Frontier Biosciences, Osaka University, Osaka 565-0871, Japan. ⁴Core Research for Evolutional Science and Technology, Japan Science and Technology Corporation, Osaka 565-0871, Japan. ⁵Sakura Motion Picture Co. Ltd, 20-1 Sendagaya 4-chome, Shibuya, Tokyo 151-0051, Japan.

nuclei were connected to the reticulocytes before this procedure, it is likely that the FACS procedure, which includes the centrifugation and suspension of cells, disconnected the extruded nuclei from the reticulocytes. In accord with a previous report⁷, the production of reticulocytes and the release of nuclei were inhibited by $10 \mu\text{g ml}^{-1}$ cytochalasin B (Fig. 2b), indicating that actin polymerization is involved in the enucleation process. Erythropoietin is indispensable for definitive erythropoiesis⁸. However, in accord with the report that the erythropoietin receptor is downregulated at early stages of erythropoiesis⁹, erythropoietin had no effect on the enucleation process (Fig. 2b). The involvement of caspase in erythroid cell differentiation has been suggested^{10,11}. But, the enucleation process described above was not prevented by caspase inhibitors (benzyloxy-carbonyl-VAD) at $100 \mu\text{M}$ (data not shown), and there was no apoptotic DNA degradation in the nuclei expelled from the erythroid precursor cells (Fig. 2c).

Apoptotic cells expose phosphatidylserine (PtdSer) on their surface as a signal for macrophages to engulf them¹². We examined whether the nuclei might also use PtdSer to be recognized by macrophages. Erythroblasts were cultured for 3 h to allow the enucleation process to begin, then stained with SYTO16, SYTOX and Cy5-labelled Annexin V and analysed by three-colour flow cytometry. As shown in Fig. 3a, nucleated erythroblasts and reticulocytes were Annexin V⁻, whereas SYTOX⁺ dead cells were Annexin V⁺. The nuclei were also Annexin V⁻. However, when the cultures

were subjected to weak physical stress and then incubated further, the nuclei became Annexin V⁺. In contrast, the reticulocytes and erythroblasts remained Annexin V⁻ after the physical stress. The Annexin V⁺ nuclei remained SYTOX⁻, indicating that the plasma membrane surrounding the nuclei remained intact after the physical stress. The exposure of PtdSer on the expelled nuclei was also observed with erythroblasts prepared from fetal liver. To further confirm the exposure of PtdSer, the erythroblasts were labelled with biotin-labelled Annexin V, followed by staining with gold-conjugated streptavidin. As shown in Fig. 3b, no gold particles were found on the nuclei protruding from reticulocytes, whereas gold labelling was observed on the outer leaflet of the continuous plasma membrane of the nuclei that were disconnected from reticulocytes. These results indicated that the nuclei exposed PtdSer only after they had been disconnected from the reticulocytes. PtdSer is normally confined to the inner leaflet of the plasma membrane¹³. This asymmetry is maintained by a combination of an ATP-dependent aminophospholipid translocase that specifically transports PtdSer from the outer to the inner leaflet of the plasma membrane, and a Ca^{2+} -dependent but energy-independent scramblase that non-specifically randomizes lipids across the bilayer. As shown in Fig. 3c, the ATP level was decreased to undetectability in the nuclei expelled from reticulocytes, whereas the Ca^{2+} level was much higher in the nuclei than in the reticulocytes (Fig. 3d, e).

Nuclei released from erythroblasts were then isolated by a cell

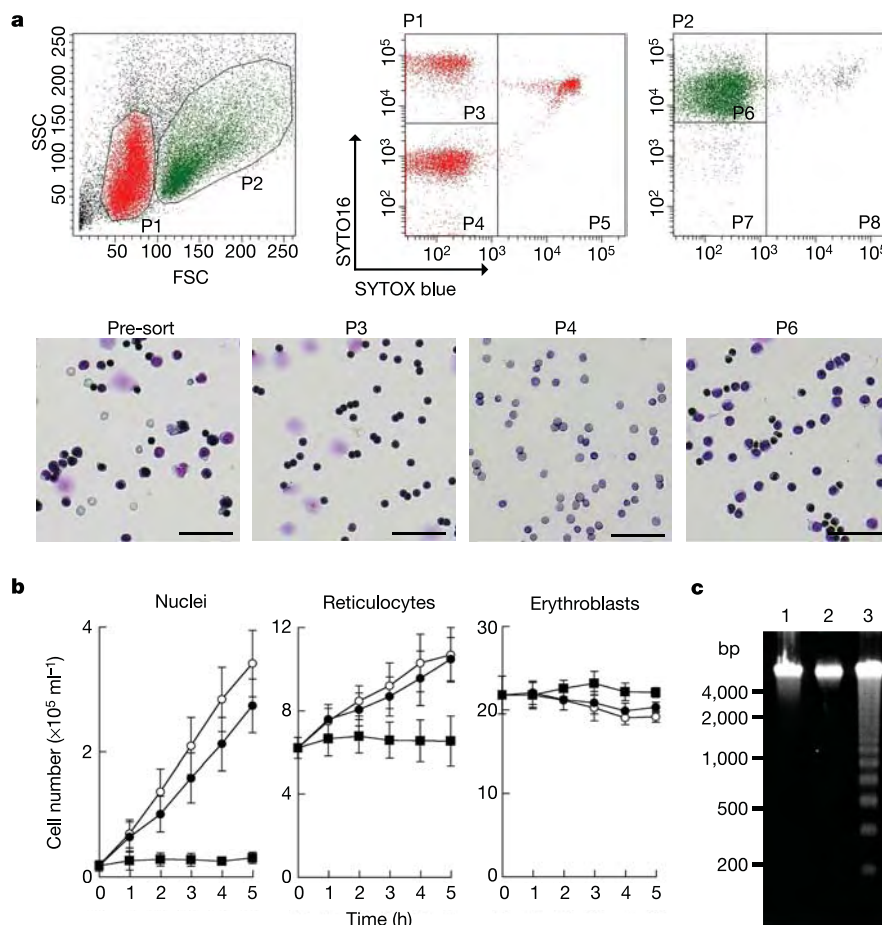


Figure 2 | Flow cytometric analysis of enucleation. **a**, Erythroblasts incubated for 5 h were stained with SYTO16 and SYTOX, then analysed by flow cytometry. Top left, a plot of side light scatter (SSC) against forward light scatter (FSC), from which events in P1 (top middle) and P2 (top right) are shown for SYTO16 and SYTOX. Bottom, cells in each fraction were stained with Wright-Giemsa. **b**, Erythroblasts were incubated with no additions (open circles), 2 units ml^{-1} erythropoietin (filled circles) or

2 units ml^{-1} erythropoietin and $10 \mu\text{g ml}^{-1}$ cytochalasin B (squares). Experiments were performed twice in triplicate; results are means $\pm$ s.d. **c**, Erythroblasts were cultured for 5 h in the absence (lane 1) or presence (lane 2) of erythropoietin. Nuclei were collected, and DNA was subjected to electrophoresis on an agarose gel. As a control, DNA from SYTOX⁺ cells (lane 3) was analysed. bp, base pairs.

sorter and used as prey for macrophages. We used macrophages from the liver of DNase II^{-/-} embryos as phagocytes to reveal the engulfed nuclei clearly¹⁴. As shown in Fig. 4a and Supplementary Movies 2 and 3, macrophages efficiently engulfed nuclei, and the phagocytosis index (the number of engulfed nuclei per macrophage) was 2.4 after incubation for 2 h (Fig. 4b). MFG-E8 binds PtdSer through its Factor VIII-homologous domains at the carboxy terminus, whereas it binds macrophages through its RGD motif at the amino terminus¹⁵. A point

mutant (D89E) of MFG-E8 in the RGD motif binds PtdSer exposed on apoptotic cells, and inhibits their engulfment by macrophages¹⁵. Similarly, the D89E mutant dose-dependently inhibited the engulfment of nuclei by macrophages, and the phagocytosis index decreased to less than 0.05 in the presence of 64 ng ml⁻¹ D89E mutant (Fig. 4b). These results indicated that PtdSer exposed on the expelled nuclei was required for their engulfment by macrophages.

Proliferating erythroblasts are closely associated with macro-

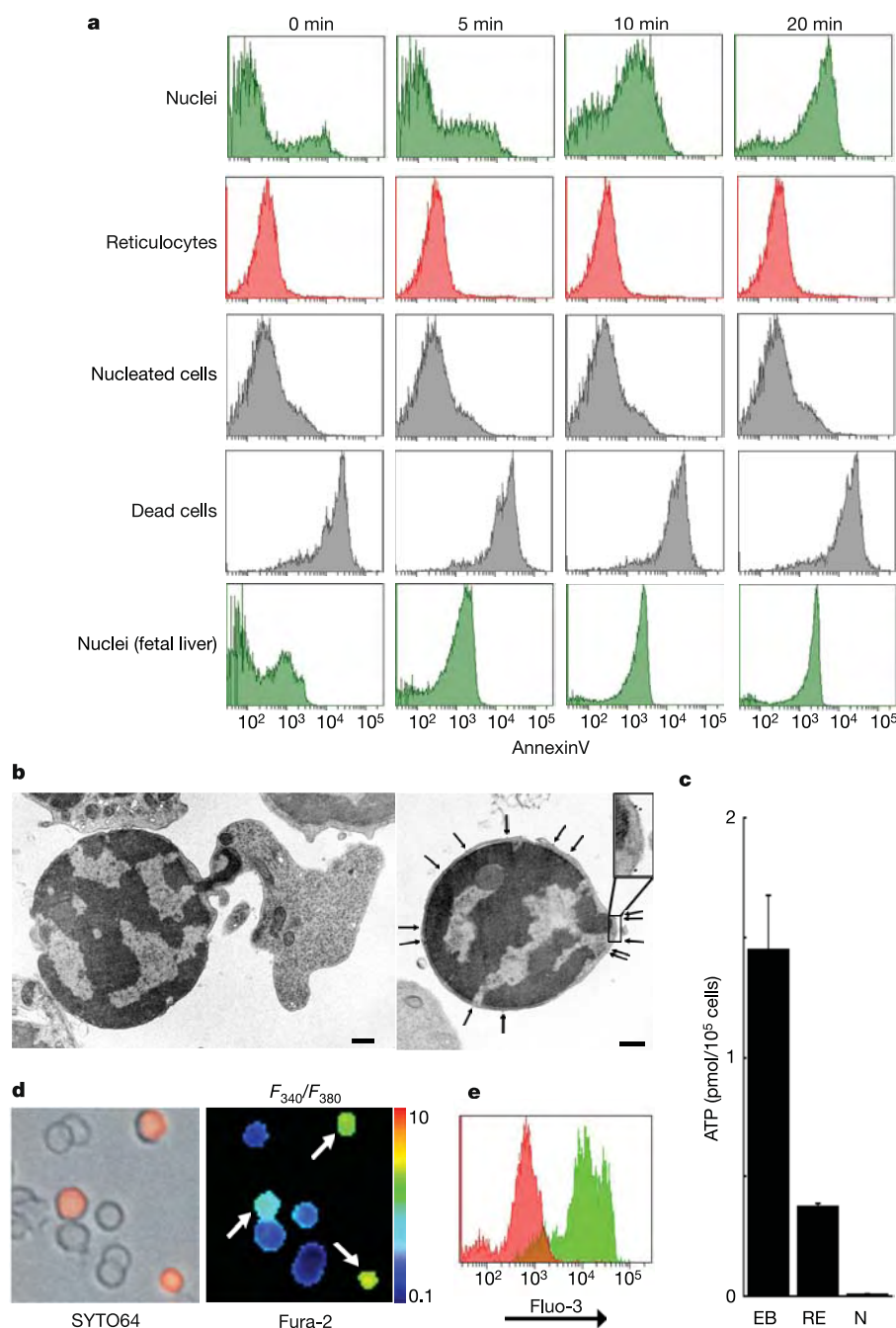


Figure 3 | Exposure of PtdSer on nuclei released from erythroblasts.

a, Erythroblasts from the spleens of phlebotomized mice were cultured for 3 h, pipetted in and out five times, and incubated further for the indicated periods. The Annexin V staining profiles for nuclei, reticulocytes, erythroblasts and dead cells are shown. The nuclei from fetal liver erythroblasts were also examined. **b**, Erythroblasts incubated for 3 h were labelled with biotin–Annexin V, followed by staining with streptavidin-conjugated colloidal gold, and analysed by electron microscopy. Scale bar, 1 μ m. Arrows indicate gold particles on plasma membranes. An enlarged

image of the boxed area is shown at top right. **c**, ATP levels for erythroblasts (EB), reticulocytes (RE) and nuclei (N). Experiments were performed twice in duplicate; results are means $\pm$ s.d. **d**, Right, erythroblasts were loaded with Fura-2. The scale at the right shows the reference image for F_{340}/F_{380} . High F_{340}/F_{380} ratios reflect high intracellular Ca^{2+} concentrations. Left, cells were stained for nuclei (red). **e**, Erythroblasts were loaded with Fluo-3 and stained with SYTO64 and SYTOX. Fluorescence-activated cell-sorting profiles for Fluo-3 in SYTO64^{low}SYTOX⁻ (reticulocytes, red) and in SYTO64^{high}SYTOX⁻ (nuclei, green) are shown.

phages in the blood islands until the nuclei are expelled, and several molecules, including cadherin, integrins and a 30-kDa protein, have been proposed to participate in the interaction between erythroblasts and macrophages¹⁶. Here we showed that the nucleus spontaneously protruded from erythroblasts, remaining connected by means of a thin membrane extension, and a weak physical force could disconnect the nucleus from the reticulocyte. Cells in blood islands are exposed to shear stress, the tangential component of haemodynamic forces¹⁷. When erythroblasts undergo enucleation, the membrane composition changes between the reticulocytes and the nuclei, with the plasma membrane surrounding the extruding nuclei becoming enriched in a receptor for concanavalin A or the binding site for merocyanine 540 (refs 18, 19). Probably as a result of such changes, the reticulocytes lose their affinity for the macrophages, whereas the nuclei remain attached to the macrophages. The exposure of erythroblasts carrying protruded nuclei to shear stress might be sufficient to release the reticulocytes physically from the nuclei and into the bloodstream.

Using the finding that protruding nuclei are disconnected from reticulocytes by weak physical stress, we showed that the expelled nuclei expose PtdSer only after they have been disconnected from the reticulocytes. The extruded nuclei contained an extremely low level of ATP. Because the nuclei contain little cytoplasm, it is likely that ATP is not supplied in the nuclei. In return, the Ca^{2+} concentration increased in the nuclei, probably as a result of inactivation of the plasma membrane Ca^{2+} ATPase, which pumps Ca^{2+} against its large concentration gradient out of the cells in an ATP-dependent manner²⁰. Thus, the aminophospholipid translocase is inactivated, whereas scramblase is activated, causing the exposure of PtdSer on the surface of nuclei. The putative aminophospholipid translocase and several scramblases have been identified^{21,22}. However, the involvement of these molecules in exposing PtdSer on the outer leaflet during apoptosis has not been proven^{23,24}. The rapid exposure of PtdSer on the nuclei expelled from erythroblasts might help in elucidating the molecular mechanism by which the asymmetrical distribution of phospholipids is maintained in the plasma membrane.

Masking PtdSer on apoptotic cells inhibits their engulfment by macrophages^{15,25,26}. Similarly, engulfment of the expelled erythroid nuclei was inhibited by the dominant-negative form of MFG-E8, indicating that PtdSer exposed on the nucleus surface works as an 'eat nucleus' signal. Several molecules expressed in macrophages have been proposed to bind PtdSer for the engulfment of apoptotic cells^{15,26,27}. Among them, MFG-E8 is expressed by thioglycollate-elicited peritoneal macrophages, immature dendritic cells and tingible-body macrophages in the spleen. However, it is not expressed in fetal liver macrophages or bone marrow macrophages²⁸. Accordingly, fetal liver macrophages from MFG-E8^{-/-} embryos engulfed the

nuclei as efficiently as wild-type fetal liver macrophages did (data not shown). It remains to be shown whether any other molecules that are proposed to participate in the engulfment of apoptotic cells are involved in the engulfment of expelled nuclei. We recently reported that the systemic administration of the dominant-negative MFG-E8 induces autoimmunity in mice by inducing the production of anti-nuclear and anti-phosphatidylserine antibodies²⁹. We thought that this was due to blockage of the engulfment of apoptotic cells. However, the adult human produces 2×10^{11} erythrocytes per day, indicating that blockage of the engulfment of nuclei might also contribute to the development of autoimmunity.

METHODS

Mice and materials. C57BL/6J mice were purchased from Charles River Japan. DNase II^{-/-} mice were described previously³. All mice were housed in a specific pathogen-free facility at Osaka University Medical School, and all animal experiments were performed in accordance with protocols approved by the Osaka University Medical School Animal Care and Use Committee.

Human erythropoietin was from Kirin Brewery and transferrin was from Sigma. The D89E mutant of mouse MFG-E8 was prepared as described¹⁵. Phycoerythrin (PE)-conjugated or fluorescein isothiocyanate (FITC)-conjugated monoclonal antibodies against Ter119, CD71, c-Kit, Thy1.2 and B220 were from BD Biosciences. Indodicarbocyanine (Cy5)-conjugated and biotin-conjugated Annexin V were from BioVision. SYTO16, SYTO64 and SYTOX were from Molecular Probes.

Preparation of erythroblasts and macrophages. Blood samples (0.5 ml) were withdrawn from ten-week-old female mice by venipuncture; four days later, erythroblasts were prepared from the mice's spleens. Erythroblasts were also prepared from fetal livers at embryonic day 13.5 (E13.5). The cell suspension was passed through a mesh, layered onto a density gradient of 70% (v/v) Percoll (Amersham Biosciences) and centrifuged at 400g for 30 min. Erythroblasts were collected and suspended in α -minimal essential medium (α -MEM) containing 10% FCS. Erythroblasts were characterized by staining with $0.25 \mu\text{g ml}^{-1}$ PE-anti-Ter119, $2.5 \mu\text{g ml}^{-1}$ FITC-anti-CD71, $1.25 \mu\text{g ml}^{-1}$ FITC-anti-Thy1.2, $0.5 \mu\text{g ml}^{-1}$ PE-anti-B220 or $5 \mu\text{g ml}^{-1}$ FITC-anti-CD117, followed by flow cytometry with a FACS Aria cell sorter (BD Biosciences).

To prepare fetal liver macrophages, livers from E14.5 DNase II^{-/-} embryos were minced, incubated at 37°C for 5 min in perfusion medium (Invitrogen) and then for 40 min in digestion medium (Invitrogen). The cells were passed through a mesh and cultured in α -MEM containing 10% FCS and mouse macrophage colony-stimulating factor³⁰. After being cultured for 2 weeks the cells expressed F4/80 and were used as fetal liver macrophages.

Enucleation assay. Erythroblasts (3×10^6 cells ml^{-1}) were incubated at 37°C in α -MEM containing 10% FCS and $500 \mu\text{g ml}^{-1}$ transferrin. After incubation, cells were stained at 22°C for 20 min with $0.25 \mu\text{M}$ SYTO16 and $0.2 \mu\text{M}$ SYTOX in 10 mM HEPES buffer pH 7.4 containing 140 mM NaCl and 5 mM MgCl_2 , and analysed by flow cytometry. To detect PtdSer, erythroblasts were stained with Cy5-labelled Annexin V in 10 mM HEPES buffer pH 7.4 containing 140 mM NaCl and 2.5 mM CaCl_2 (binding buffer). After being washed with binding buffer, the cells were further stained with SYTO16 and SYTOX, then analysed by flow cytometry. For cytochemical analysis the cells were cytospun, then stained with Wright-Giemsa.

Assay for intracellular ATP and Ca^{2+} . Erythroblasts (3×10^7 cells) from the spleens of phlebotomized mice were incubated at 37°C for 3 h, and FSC^{high} SYTO16^{high} SYTOX⁻ erythroblasts, FSC^{low} SYTO16^{high} SYTOX⁻ nuclei and FSC^{low} SYTO16^{low/-} SYTOX⁻ reticulocytes were collected with a FACS Aria cell sorter. The ATP level was then determined with an ATP Bioluminescence Assay Kit CLSII (Roche) using 10^5 cells or nuclei. For determining the cellular Ca^{2+} level, erythroblasts were treated at 37°C for 30 min with $1.8 \mu\text{M}$ Fluo-3 acetoxymethyl ester or $1 \mu\text{M}$ Fura-2 acetoxymethyl ester (Dojindo) in α -MEM containing 0.02% Pluronic F-127 (Molecular Probes). The cells loaded with Fluo-3 were further stained with $2 \mu\text{M}$ SYTO64 and $0.2 \mu\text{M}$ SYTOX, then analysed by flow cytometry. The Fura-2-loaded cells were irradiated with ultraviolet at 340 nm followed by ultraviolet at 380 nm. Emission fluorescence, F , was recorded on an ORCA-ER charge-coupled device camera (Hamamatsu Photonics) through a Diaphoto 300 fluorescence microscope (Nikon), and the data processed with an AquaCosmos image processor (Hamamatsu Photonics) to obtain the ratio F_{340}/F_{380} .

Engulfment of nuclei by macrophages. To prepare nuclei, 3×10^7 erythroblasts were incubated at 37°C for 5 h and stained with SYTO16 and SYTOX. The nuclei, represented by FSC^{low} SYTO16^{high} SYTOX⁻, were collected with a cell sorter. DNase II^{-/-} fetal liver macrophages (2×10^4 cells) were cultured in

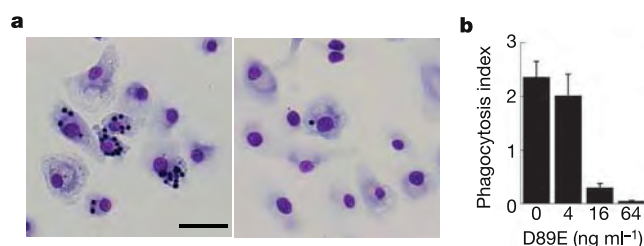


Figure 4 | Phosphatidylserine-dependent engulfment of nuclei. **a**, Fetal liver macrophages were co-cultured with nuclei in the absence (left) or presence (right) of 64 ng ml^{-1} D89E. The cells were then stained with Wright-Giemsa. Scale bar, $50 \mu\text{m}$. **b**, Engulfment of nuclei by fetal liver macrophages was performed in the presence of the indicated concentrations of D89E. The experiments were performed three times; results are values of the phagocytosis index (the number of engulfed nuclei per macrophage) and are shown as means $\pm$ s.d.

α -MEM containing 10% FCS on eight-well Lab-Tek II chamber slides (Nalge Nunc). Nuclei (4×10^5) were added to the macrophages and incubated at 37 °C for 2 h. Unengulfed nuclei were removed by vigorous washing, and the macrophages were stained with Wright-Giemsa.

Electron microscopy. Cells were collected by centrifugation at 500g for 5 min, washed with PBS, fixed with 2% glutaraldehyde, 2% paraformaldehyde in 0.1 M phosphate buffer pH 7.2, postfixed with 1% OsO_4 at 4 °C for 2 h and embedded in Epon 812 as described³. To detect the binding of Annexin V to PtdSer at the ultrastructural level, erythroblasts were incubated with biotin-labelled Annexin V as described above, then fixed with 0.2% glutaraldehyde, 4% paraformaldehyde. After being rinsed with PBS, they were incubated with streptavidin-conjugated 15-nm colloidal gold (British Biocell International), fixed again with 2% glutaraldehyde, 2% paraformaldehyde in 0.1 M phosphate buffer pH 7.2, and embedded in Epon 812. Sections (80 nm) were prepared with a Reichert Ultracut N ultramicrotome (Nissei), stained with lead citrate and uranyl acetate, and observed with a Hitachi H-7100 electron microscope.

Received 29 May; accepted 29 June 2005.

- Bessis, M., Mize, C. & Prenant, M. Erythropoiesis: comparison of *in vivo* and *in vitro* amplification. *Blood Cells* **4**, 155–174 (1978).
- Sadahira, Y. & Mori, M. Role of the macrophage in erythropoiesis. *Pathol. Int.* **49**, 841–848 (1999).
- Kawane, K. *et al.* Requirement of DNase II for definitive erythropoiesis in the mouse fetal liver. *Science* **292**, 1546–1549 (2001).
- Pannacchiulli, I. M. *et al.* Effect of bleeding on *in vivo* and *in vitro* colony-forming hemopoietic cells. *Acta Haematol.* **58**, 27–33 (1977).
- Zhang, D. *et al.* An optimized system for studies of EPO-dependent murine pro-erythroblast development. *Exp. Hematol.* **29**, 1278–1288 (2001).
- Socolovsky, M. *et al.* Ineffective erythropoiesis in Stat5a^{-/-}5b^{-/-} mice due to decreased survival of early erythroblasts. *Blood* **98**, 3261–3273 (2001).
- Repasky, E. A. & Eckert, B. S. The effect of cytochalasin B on the enucleation of erythroid cells *in vitro*. *Cell Tissue Res.* **221**, 85–91 (1981).
- Wu, H., Liu, X., Jaenisch, R. & Lodish, H. F. Generation of committed erythroid BFU-E and CFU-E progenitors does not require erythropoietin or the erythropoietin receptor. *Cell* **83**, 59–67 (1995).
- Broudy, V. C., Lin, N., Brice, M., Nakamoto, B. & Papayannopoulou, T. Erythropoietin receptor characteristics on primary human erythroid cells. *Blood* **77**, 2583–2590 (1991).
- Carlisle, G. W., Smith, D. H. & Wiedmann, M. Caspase-3 has a nonapoptotic function in erythroid maturation. *Blood* **103**, 4310–4316 (2004).
- Zermati, Y. *et al.* Caspase activation is required for terminal erythroid differentiation. *J. Exp. Med.* **193**, 247–254 (2001).
- Fadok, V. A., Bratton, D. L., Frasch, S. C., Warner, M. L. & Henson, P. M. The role of phosphatidylserine in recognition of apoptotic cells by phagocytes. *Cell Death Differ.* **5**, 551–562 (1998).
- Balasubramanian, K. & Schroit, A. J. Aminophospholipid asymmetry: A matter of life and death. *Annu. Rev. Physiol.* **65**, 701–734 (2003).
- Kawane, K. *et al.* Impaired thymic development in mouse embryos deficient in apoptotic DNA degradation. *Nature Immunol.* **4**, 138–144 (2003).
- Hanayama, R. *et al.* Identification of a factor that links apoptotic cells to phagocytes. *Nature* **417**, 182–187 (2002).
- Hanspal, M. Importance of cell–cell interactions in regulation of erythropoiesis. *Curr. Opin. Hematol.* **4**, 142–147 (1997).
- Davies, P. F. Flow-mediated endothelial mechanotransduction. *Physiol. Rev.* **75**, 519–560 (1995).
- Schlegel, R. A., Phelps, B. M., Waggoner, A., Terada, L. & Williamson, P. Binding of merocyanine 540 to normal and leukemic erythroid cells. *Cell* **20**, 321–328 (1980).
- Geiduschek, J. B. & Singer, S. J. Molecular changes in the membranes of mouse erythroid cells accompanying differentiation. *Cell* **16**, 149–163 (1979).
- Strehler, E. E. & Treiman, M. Calcium pumps of plasma membrane and cell interior. *Curr. Mol. Med.* **4**, 323–335 (2004).
- Tang, X., Halleck, M. S., Schlegel, R. A. & Williamson, P. A subfamily of P-type ATPases with aminophospholipid transporting activity. *Science* **272**, 1495–1497 (1996).
- Zhou, Q. *et al.* Molecular cloning of human plasma membrane phospholipid scramblase. A protein mediating transbilayer movement of plasma membrane phospholipids. *J. Biol. Chem.* **272**, 18240–18244 (1997).
- Zhou, Q., Zhao, J., Wiedmer, T. & Sims, P. J. Normal hemostasis but defective hematopoietic response to growth factors in mice deficient in phospholipid scramblase 1. *Blood* **99**, 4030–4038 (2002).
- Williamson, P. & Schlegel, R. A. Hide and seek: the secret identity of the phosphatidylserine receptor. *J. Biol. Chem.* **279**, 14 (2004).
- Krahling, S., Callahan, M. K., Williamson, P. & Schlegel, R. A. Exposure of phosphatidylserine is a general feature in the phagocytosis of apoptotic lymphocytes by macrophages. *Cell Death Differ.* **6**, 183–189 (1999).
- Fadok, V. A. *et al.* A receptor for phosphatidylserine-specific clearance of apoptotic cells. *Nature* **405**, 85–90 (2000).
- Kawasaki, Y., Nakagawa, A., Nagaosa, K., Shiratsuchi, A. & Nakanishi, Y. Phosphatidylserine binding of class B scavenger receptor type I, a phagocytosis receptor of testicular Sertoli cells. *J. Biol. Chem.* **277**, 27559–27566 (2002).
- Hanayama, R., Tanaka, M., Miwa, K. & Nagata, S. Expression of developmental endothelial locus-1 in a subset of macrophages for engulfment of apoptotic cells. *J. Immunol.* **172**, 3876–3882 (2004).
- Asano, K. *et al.* Masking of phosphatidylserine inhibits apoptotic cell engulfment and induces autoantibody production in mice. *J. Exp. Med.* **200**, 459–467 (2004).
- Takeshita, S., Kaji, K. & Kudo, A. Identification and characterization of the new osteoclast progenitor with macrophage phenotypes being able to differentiate into mature osteoclasts. *J. Bone Miner. Res.* **15**, 1477–1488 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Fukuyama for help at the initial stage of this work; H. Suzuki, I. Ito and S. Hanasaki for taking the time-lapse video; T. Yanagida for discussion; and M. Fujii and M. Harayama for secretarial assistance. This work was supported in part by Grants-in-Aid from the Ministry of Education, Science, Sports and Culture in Japan.

Author Contributions H.Y. performed most of the experiments. M.K. carried out the electron microscopy analysis, and Y.M. took the time-lapse video. K.K., Y.U. and S.N. contributed to the data analysis. S.N. was responsible for planning the project.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.N. (Nagata@genetic.med.osaka-u.ac.jp).

A SUMOylation-dependent pathway mediates transrepression of inflammatory response genes by PPAR- γ

Gabriel Pascual^{1,4}, Amy L. Fong^{1,4}, Sumito Ogawa¹, Amir Gamliel², Andrew C. Li¹, Valentina Perissi², David W. Rose³, Timothy M. Willson⁵, Michael G. Rosenfeld^{2,3} & Christopher K. Glass^{1,3}

Peroxisome proliferator-activated receptor- γ (PPAR- γ) has essential roles in adipogenesis and glucose homeostasis, and is a molecular target of insulin-sensitizing drugs^{1–3}. Although the ability of PPAR- γ agonists to antagonize inflammatory responses by transrepression of nuclear factor kappa B (NF- κ B) target genes is linked to antidiabetic⁴ and antiatherogenic actions⁵, the mechanisms remain poorly understood. Here we report the identification of a molecular pathway by which PPAR- γ represses the transcriptional activation of inflammatory response genes in mouse macrophages. The initial step of this pathway involves ligand-dependent SUMOylation of the PPAR- γ ligand-binding domain, which targets PPAR- γ to nuclear receptor corepressor (NCoR)–histone deacetylase-3 (HDAC3) complexes on inflammatory gene promoters. This in turn prevents recruitment of the ubiquitylation/19S proteasome machinery that normally mediates the signal-dependent removal of corepressor complexes required for gene activation. As a result, NCoR complexes are not cleared from the promoter and target genes are maintained in a repressed state. This mechanism provides an explanation for how an agonist-bound nuclear receptor can be converted from an activator of transcription to a promoter-specific repressor of NF- κ B target genes that regulate immunity and homeostasis.

NCoR and the related factor SMRT (silencing mediator of retinoic acid and thyroid hormone receptors) are components of corepressor complexes containing HDAC3, transducin beta-like protein-1 (TBL1) and TBLR1 that interact with a subset of unliganded nuclear receptors and mediate active transcriptional repression^{6–12}. Ligand-dependent dismissal of these complexes requires Ubc5-dependent ubiquitylation and proteasomal degradation, with TBLR1 functioning as an essential E3 ligase¹³. Recent studies indicate that NCoR/SMRT complexes are also required for basal repression of a subset of NF- κ B and AP-1 target genes^{13–15}, with loss of NCoR resulting in a partially activated phenotype in macrophages¹⁴. We note that a number of inflammatory response genes that are derepressed in NCoR-deficient macrophages are also subject to transrepression by PPAR- γ agonists, suggesting a possible role for NCoR in this process.

We focused on the mouse inducible nitric oxide synthase gene (*iNOS*, also known as *Nos2*) as a model because it is one of several inflammatory response genes expressed by macrophages that is strongly induced by lipopolysaccharide (LPS)¹⁶ and negatively regulated by PPAR- γ agonists¹⁷. Inhibition of NCoR expression using an NCoR-specific short interfering (si)RNA validated for efficacy¹³ (Supplementary Fig. S1) resulted in complete reversal of *iNOS* transrepression by the synthetic PPAR- γ ligands rosiglitazone and GW0072 (ref. 18) (Fig. 1a). Consistent with these findings, knock-

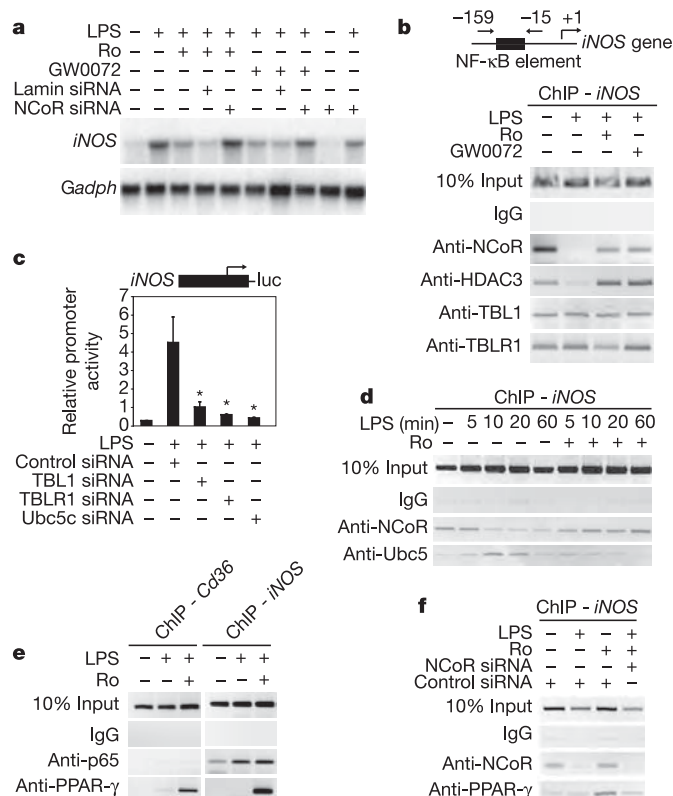


Figure 1 | PPAR- γ prevents LPS-induced dissociation of the NCoR-HDAC3 complex from the *iNOS* promoter. **a**, Northern blot analysis indicating that siRNAs directed against NCoR abolish rosiglitazone- (Ro) and GW0072-dependent repression of LPS-induced *iNOS* expression. Lamin siRNA was used as a control **b**, Rosiglitazone and GW0072 inhibit release of NCoR from the *iNOS* promoter as shown by ChIP assay. The fraction of cell lysate used in the immunoprecipitation is shown (10% input). **c**, siRNAs directed against TBL1, TBLR1 or Ubc5c prevent LPS induction of the *iNOS* promoter in transiently transfected RAW264.7 macrophages. Error bars represent standard deviations of triplicate, independent determinations. Asterisks, $P < 0.01$ compared to LPS induction. **d**, Rosiglitazone prevents LPS-dependent recruitment of Ubc5 to the *iNOS* promoter as detected by ChIP assay. **e**, PPAR- γ binds to the *Cd36* and *iNOS* promoters in a ligand-dependent manner as detected by ChIP assay. **f**, Knockdown of NCoR expression prevents PPAR- γ recruitment to the *iNOS* promoter as detected by ChIP assay.

¹Department of Cellular and Molecular Medicine, ²Howard Hughes Medical Institute, ³Department of Medicine, and ⁴Biomedical Sciences Graduate Program, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. ⁵GlaxoSmithKline, 5 Moore Drive, PO Box 13398, Research Triangle Park, North Carolina 27709, USA.

down of NCoR expression, but not SMRT expression, resulted in the reversal of repression of an *iNOS* promoter reporter by PPAR- γ in RAW264.7 macrophages (Supplementary Fig. S2a). Potential roles for NCoR-associated HDAC proteins were supported by the finding that treatment with 10 nM of the histone deacetylase inhibitor trichostatin A reversed rosiglitazone-dependent transrepression of *iNOS* (Supplementary Fig. S2b). To evaluate specifically the role of HDAC3, the effect of a validated HDAC3-specific pool of siRNAs (Supplementary Fig. S1) was tested in RAW264.7 cells. The HDAC3-specific siRNAs, but not control siRNAs directed against HDAC7, reversed the transrepression observed on the *iNOS* promoter in this system (Supplementary Fig. S2c).

These observations predicted that NCoR–HDAC3–TBL complexes should associate with the *iNOS* promoter. Chromatin immunoprecipitation (ChIP) experiments confirmed that NCoR, HDAC3, TBL1 and TBLR1 were present on the *iNOS* promoter under basal conditions, and that the NCoR and HDAC3 components cleared following LPS stimulation (Fig. 1b). However, in cells treated with rosiglitazone or GW0072, both NCoR and HDAC3 remained on the *iNOS* promoter after LPS stimulation (Fig. 1b). Because signal-dependent induction of NF- κ B target genes has been suggested to require removal of NCoR complexes through Ubc5-dependent ubiquitylation¹³, we used validated siRNAs directed against TBL1, TBLR1 and the ubiquitin conjugating enzyme Ubc5c (ref. 13, also known as Ube2d3), all of which inhibited *iNOS* induction in response to LPS (Fig. 1c). These data are consistent with the hypothesis that the TBL1–TBLR1-recruited ubiquitylation complex is required for LPS-dependent clearance of NCoR and HDAC3 from the *iNOS* promoter, and that PPAR- γ represses *iNOS* activation by preventing TBL1–TBLR1-dependent corepressor clearance.

We next used ChIP assays to evaluate whether an ordered sequence of events was required for NCoR clearance in response to LPS stimulation in the presence or absence of rosiglitazone. Although NCoR was cleared from the *iNOS* promoter within 10 min of LPS treatment, pretreatment of cells with rosiglitazone inhibited clearance at all time points tested (Fig. 1d). Notably, Ubc5 was rapidly recruited to the *iNOS* promoter following LPS stimulation in the absence of rosiglitazone, but was not recruited to the promoter when rosiglitazone was present (Fig. 1d). These results suggest that PPAR- γ acts to repress LPS induction of the *iNOS* gene by preventing recruitment of the Ubc5/19S proteasome machinery required for the clearance of NCoR and HDAC3.

Evaluation of the *iNOS* promoter (negatively regulated by PPAR- γ) and the *Cd36* promoter (positively regulated by PPAR- γ)¹⁹ in macrophages revealed that PPAR- γ was recruited to both promoters in a ligand-dependent manner (Fig. 1e). As expected, the p65 component of NF- κ B was recruited exclusively to the *iNOS* promoter in response to LPS, and was not affected by rosiglitazone treatment (Fig. 1e). The recruitment of PPAR- γ to the *iNOS* promoter did not involve sequence-specific DNA binding because a PPAR- γ mutant (PPAR- γ ^{C126A/E127A}) containing amino acid substitutions in the DNA-binding domain that abolish binding to PPAR- γ response elements and its recruitment to the *Cd36* promoter, was efficiently recruited to the *iNOS* promoter (Supplementary Fig. S3a). This is consistent with previous studies of PPAR- γ ^{C126A/E127A} indicating that it does not activate PPAR- γ target genes but that it retains transrepression activity²⁰. Ligand-dependent interaction of PPAR- γ with the *iNOS* promoter was abolished by siRNA-mediated knockdown of NCoR, indicating that NCoR is required for PPAR- γ recruitment (Fig. 1f). Similar results were obtained for four additional LPS-inducible, PPAR- γ -sensitive promoters: *Ccl3*, *Ccl7*, *Cxcl10* and *Tgtp*¹⁴ (Supplementary Fig. S3b, c).

The observation that ligand-dependent recruitment of PPAR- γ to LPS-responsive promoters requires NCoR raises a paradox, because ligand binding disrupts direct interactions between NCoR and PPAR- γ ¹⁸. To identify PPAR- γ -interacting proteins that might potentially resolve this paradox, a yeast two-hybrid screen was performed

using a library constructed from messenger RNA derived from primary mouse macrophages. One of the clones isolated in this screen encoded the first 208 amino acids of PIAS1 (protein inhibitor of activated STAT1), initially identified as a suppressor of interferon-dependent transcription²¹ and now known to belong to a family of SUMO E3 ligases²². The region of PIAS1 isolated in this screen (hereafter referred to as PIAS1-N, see Supplementary Fig. S4a) contains motifs previously shown to interact with various nuclear receptors^{23–25}. The interaction between PPAR- γ and the PIAS1 clone was confirmed both by yeast survival and α -galactosidase liquid assays (Supplementary Fig. S4b). Furthermore, co-immunoprecipitation experiments using antibodies directed against either endogenous or epitope-tagged PPAR- γ and PIAS1 showed a basal interaction in RAW264.7 cells and primary macrophages that was modestly enhanced by treatment with rosiglitazone (Fig. 2a).

SUMOylation of transcription factors has previously been correlated with impaired transcriptional activation and/or transcriptional repression^{26–28}. The PIAS1-N fragment dominantly inhibited PPAR- γ -dependent transrepression of the *iNOS* promoter in RAW264.7 cells (Supplementary Fig. S4c), suggesting a role for PIAS1 or other PIAS proteins in this process. To specifically evaluate the consequences of loss of PIAS1 expression, transfection of validated PIAS1 siRNAs (Supplementary Fig. S1) resulted in significant inhibition of PPAR- γ -dependent repression of the *iNOS* promoter, similar to the effects observed for NCoR siRNA (Fig. 2b), but did not impair transcriptional activation of a positively regulated PPAR- γ

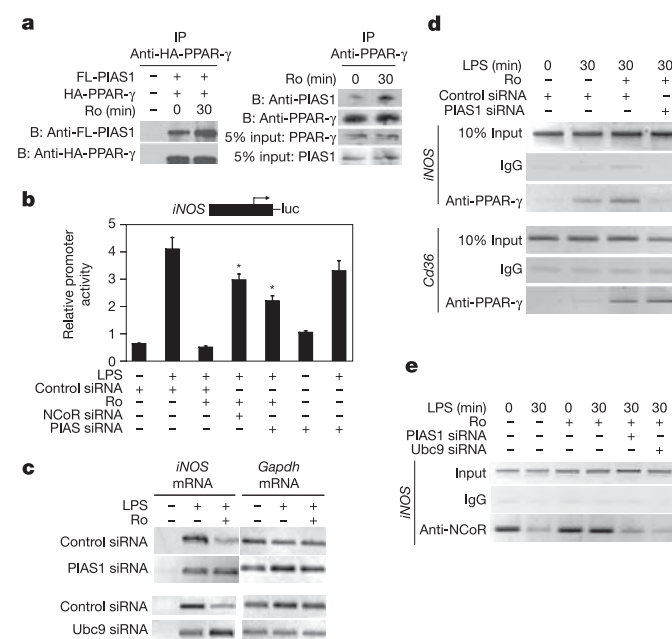


Figure 2 | PIAS1 interacts with PPAR- γ and is required for transrepression of *iNOS*. **a**, Immunoblots (B) for FLAG (FL)-tagged PIAS1 and HA-tagged PPAR- γ in transfected RAW264.7 macrophages (left panel) or endogenous PIAS1 and PPAR- γ in primary macrophages (right panel) indicate PIAS1 interaction with immunoprecipitated PPAR- γ . **b**, siRNAs directed against NCoR and PIAS1 in RAW264.7 macrophages reverse PPAR- γ transrepression of the *iNOS* promoter. Error bars represent standard deviations of triplicate, independent determinations. Asterisk, $P < 0.001$ compared to rosiglitazone repression. **c**, Primary macrophages transfected with control siRNA, PIAS1- or Ubc9-specific siRNA show reversal of PPAR- γ transrepression by semiquantitative PCR for endogenous *Gapdh* and *iNOS* expression. **d**, PPAR- γ recruitment to the *iNOS* but not the *Cd36* promoter in primary macrophages requires PIAS1, as detected by ChIP assays. **e**, Prevention of signal-dependent dissociation of NCoR from the *iNOS* promoter by rosiglitazone requires PIAS1 and Ubc9, as detected by ChIP assays.

target gene (data not shown). Moreover, siRNA-mediated knockdown of PIAS1 in primary macrophages abolished PPAR- γ transrepression of the endogenous *iNOS* gene (Fig. 2c). Finally, knockdown of Ubc9, the SUMOylation pathway rate-limiting E2 ligase, significantly impaired PPAR- γ -dependent transrepression of *iNOS* in both RAW264.7 cells and primary macrophages (Fig. 2c and Supplementary Fig. S5a). These results suggest that PIAS1/Ubc9-mediated PPAR- γ SUMOylation is required for PPAR- γ -dependent transrepression.

ChIP assays were next performed in macrophages to determine the roles of PIAS1 and Ubc9 in ligand-dependent recruitment of PPAR- γ to the *iNOS* promoter and in the prevention of NCoR clearance. Knockdown of PIAS1 expression abolished recruitment of PPAR- γ to the *iNOS* promoter, but did not affect recruitment to the positively regulated *Cd36* promoter (Fig. 2d). In addition, knockdown of PIAS1 or Ubc9 prevented the retention of NCoR on the *iNOS* promoter in the presence of rosiglitazone and LPS (Fig. 2e).

SUMOylation of the PPAR- γ 2 AF-1 domain at residue K107 (equivalent to K77 of PPAR- γ 1) inhibits ligand-dependent activation of positively regulated target genes²⁹. The primary amino acid sequence of murine PPAR- γ shows an additional SUMOylation consensus sequence (ψ KXE/D; where ψ represents a hydrophobic amino acid and X refers to any amino acid)³⁰ at K365 (corresponding to K367 in the human PPAR- γ sequence, Fig. 3a). Notably, crystal structures of the apo and rosiglitazone-bound forms of PPAR- γ indicate that the primary amine group of K365 is oriented towards the interior of the ligand-binding domain in the apo form, but is solvent-exposed in the rosiglitazone-bound form (Fig. 3a). Because this amino group is the point of covalent attachment of SUMO, the PPAR- γ crystal structures suggest that K365 could be SUMOylated in a ligand-dependent manner. To test this hypothesis, K365 of PPAR- γ was mutated to arginine (K365R), and the wild-type and mutant proteins were tested for SUMOylation *in vivo* and *in vitro*. Wild-type PPAR- γ , but not PPAR- γ ^{K365R}, showed a significant increase in SUMOylation following treatment with rosiglitazone (Fig. 3b and data not shown).

To determine the functional consequences of K77- and K365-dependent SUMOylation, the ability of each mutant to inhibit the *iNOS* promoter and/or transactivate a positively regulated PPAR- γ -dependent promoter was tested in RAW264.7 cells. PPAR- γ ^{K365R} did not inhibit the *iNOS* promoter, whereas PPAR- γ ^{K77R} retained full transrepression activity (Fig. 3c). PPAR- γ ^{K77R} showed enhanced transactivation function, consistent with previous findings²⁹, whereas PPAR- γ ^{K365R} showed approximately the same activity as wild-type PPAR- γ on the positively regulated Aox-TK luciferase promoter (Fig. 3d). ChIP assays indicated that wild-type PPAR- γ and PPAR- γ ^{K77R} were efficiently recruited to the *iNOS* promoter in response to rosiglitazone, whereas PPAR- γ ^{K365R} was not (Fig. 3e). In contrast, wild-type PPAR- γ and each of the PPAR- γ mutants were recruited to the positively regulated *Cd36* promoter (Fig. 3f).

Mammalian two-hybrid assays were used to explore effects of SUMOylation on interactions of PPAR- γ with NCoR and HDAC3. Previous studies have demonstrated that unliganded PPAR- γ binds to one of two nuclear receptor interaction domains in the extreme C terminus of NCoR (termed the IDC), and that this interaction is reversed by ligand¹⁸. Mammalian two-hybrid assays confirmed this interaction and demonstrated that knocking down expression of PIAS1 or Ubc9 did not influence ligand-dependent dissociation of PPAR- γ from the IDC (Fig. 4a). In contrast, PPAR- γ showed a ligand-dependent increase in interaction with an NCoR deletion mutant lacking the IDC that was abolished by knocking down Ubc9 or PIAS1 (Fig. 4a and Supplementary Fig. S5b). Mammalian two-hybrid assays also showed an interaction between PPAR- γ and HDAC3, but this interaction was only modestly affected by ligand or by knockdown of PIAS1 or Ubc9 (Supplementary Fig. S5c). Consistent with these results, knockdown of HDAC3 reduced, but did not prevent, recruitment of PPAR- γ to the *iNOS* promoter in

response to ligand, as determined by ChIP assays (Fig. 4b). These results suggest that NCoR is required for ligand-dependent recruitment of SUMOylated PPAR- γ to the *iNOS* promoter and that HDAC3 plays a quantitative role in stabilizing this interaction.

The present studies define sequential steps of a pathway mediating ligand-dependent transrepression of inflammatory response genes by PPAR- γ in macrophages (Fig. 4c). Genes subject to transrepression by this pathway are marked in the basal state by the presence of NCoR–HDAC3–TBL corepressor complexes. LPS signalling results in the clearance of the NCoR and HDAC3 components of this complex in a TBL1-, TBLR1- and Ubc5-dependent manner, allowing a switch from active repression to transcriptional activation. The PPAR- γ -dependent transrepression pathway is initiated by ligand-induced SUMOylation

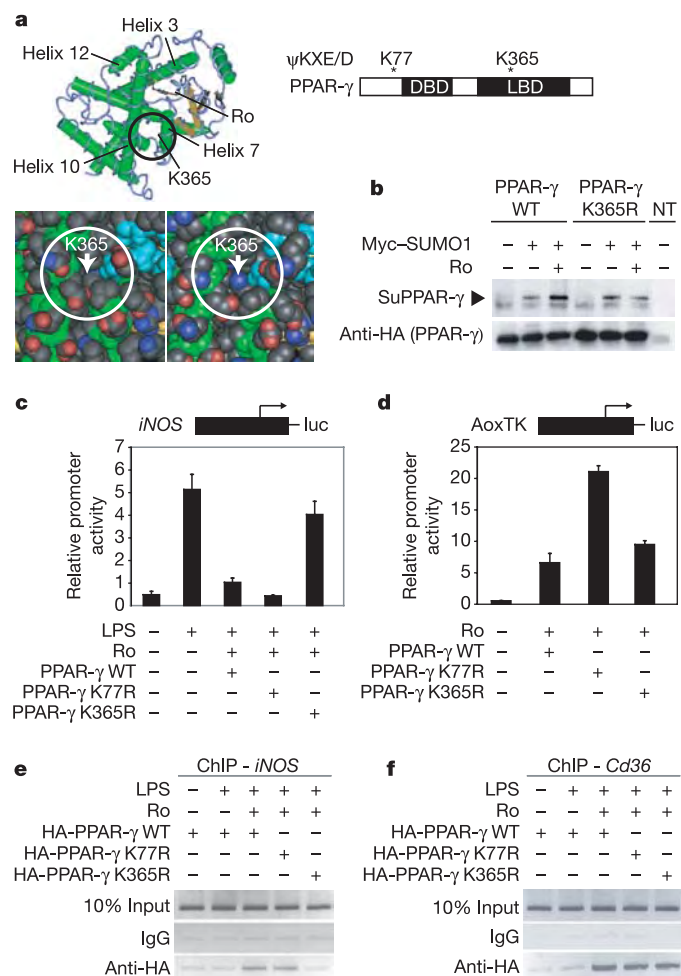


Figure 3 | Ligand-dependent SUMOylation of PPAR- γ is required for transrepression. **a**, Top right, a schematic representation of PPAR- γ with two consensus SUMOylation sites at K77 and K365. Top left, the PPAR- γ ligand-binding domain (LBD), indicating the location of K365 in helix 7. Bottom panels, close-up views of this region of the apo PPAR- γ LBD surface, with the arrow pointing to a hydrogen atom in the K365 side chain (left), and rosiglitazone-bound PPAR- γ LBD (right), with the arrow pointing to solvent-exposure of the nitrogen atom of the primary amine of K365 (shown in blue). **b**, SUMOylation of PPAR- γ at K365 occurs in a ligand-dependent manner. WT, wild type; NT, non-transfected cells; SuPPAR- γ , SUMOylated PPAR- γ species. HA-tagged WT or K365R PPAR- γ was immunoprecipitated from cell lysates and immunoblotted for HA-tag or SUMO Myc-tag. **c**, **d**, SUMOylation of PPAR- γ at K365 but not K77 is required for transrepression of the *iNOS* promoter (**c**), but not transactivation of the AoxTK promoter (**d**), in transfected RAW264.7 macrophages. Error bars represent standard deviations of triplicate, independent determinations. **e**, **f**, SUMOylation of PPAR- γ at K365 is required for recruitment to the *iNOS* promoter (**e**) but not the *Cd36* promoter (**f**) in RAW264.7 cells, as demonstrated by ChIP assays.

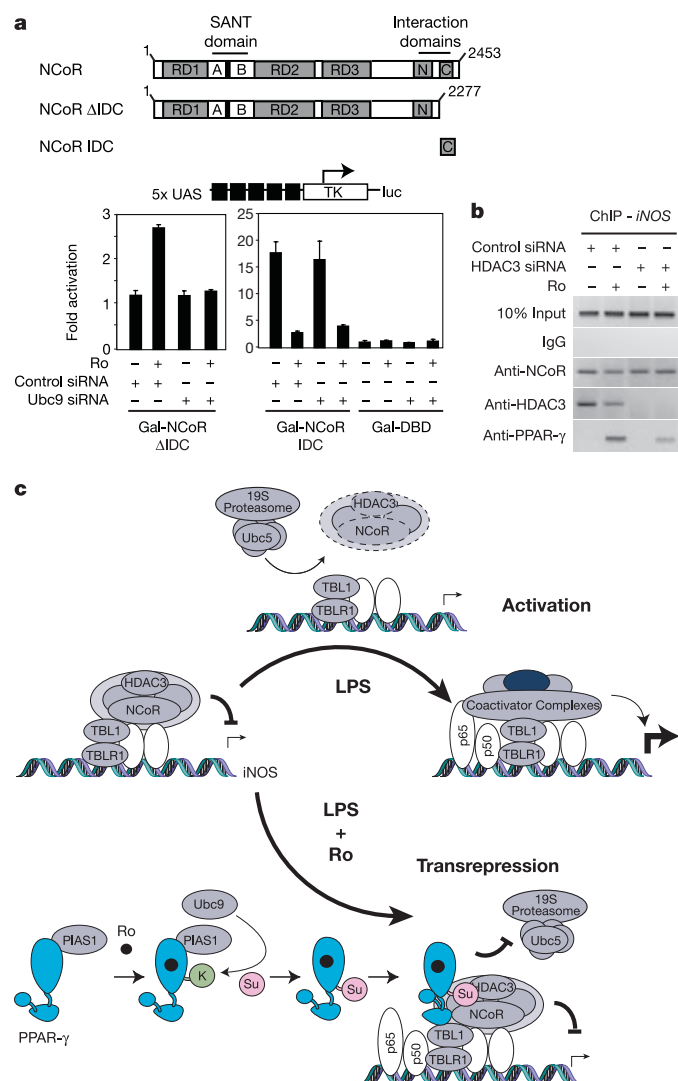


Figure 4 | SUMOylation of PPAR- γ promotes interaction with the NCoR-HDAC3 complex. **a**, Mammalian two-hybrid assay in RAW264.7 cells, indicating that wild-type VP16-PPAR- γ interacts with Gal-DBD-NCOR ΔIDC (NCOR amino acids 1–2277, without the IDC), but not Gal-DBD-NCOR IDC (the C-terminal nuclear receptor interaction motif of NCOR), in a ligand-dependent and Ubc9-dependent manner. **b**, siRNA directed against HDAC3 reduces but does not abolish recruitment of PPAR- γ to the iNOS promoter in primary macrophages, as demonstrated by ChIP assays. **c**, Model for mechanisms of LPS activation and PPAR- γ -dependent transrepression of the iNOS gene. See text for details. K, SUMO target lysine within PPAR- γ DNA-binding domain; Su, SUMO conjugate on target lysine.

of the ligand-binding domain. This modification targets PPAR- γ to NCoR complexes associated with the promoter, preventing Ubc5 recruitment in response to LPS signals. As a result, NCoR complexes are not cleared from the promoter and target genes are maintained in a repressed state. Interestingly, allosteric changes in the PPAR- γ ligand-binding domain required for entry into the SUMOylation-dependent transrepression pathway are distinct from changes that regulate interactions with conventional coregulators. It will be of interest to define the extent to which this pathway is used by PPAR- γ and other nuclear receptors, and to explore how this mechanism can be exploited to develop new drugs for the treatment of inflammatory and metabolic diseases.

METHODS

Plasmids and cell culture. Primary macrophages were elicited by intraperitoneal injection with 2 ml of thioglycollate. HA-tagged wild-type and K77R and

K365R PPAR- γ mutants were cloned into a pcDNA3 backbone (Invitrogen). Wild-type (WT)-PIAS1 and PIAS1-N were cloned into a 2 $\times$ FLAG-pcDNA3 expression vector. PPAR- γ bait used for the yeast two-hybrid, including the DNA-binding domain, hinge region and ligand-binding domain (PPAR- γ DHL), was inserted into the pGBK7 vector (Clontech). For RNAi experiments, smart-pool siRNAs (Dharmacon) against PIAS1, Ubc9, HDAC3, HDAC7 or control non-specific and (previously validated) NCoR were transfected into primary macrophages using lipofectamine 2000 (Invitrogen) and incubated for 48 h. Effects of these siRNAs on cellular protein levels are illustrated in Supplementary Fig. S1.

Yeast two-hybrid screen. A yeast two-hybrid library was generated with the pGAD vector (Clontech) with RNA derived from primary peritoneal macrophages elicited from normal and hypercholesterolaemic mice. The library was transformed into the AH109 yeast strain and was mated to Y187 strain transformed with PPAR- γ DHL bait. Colonies were picked 4–6 days after mating. Polymerase chain reaction (PCR) inserts were amplified and sequenced. Yeast plasmids were purified from individual clones. Finally, interactions were verified by α -galactosidase activity in yeast liquid culture assays.

Transient transfection. The RAW264.7 mouse macrophage cell line was transiently transfected with iNOS or Aox-TK promoters directing luciferase expression, as previously described¹⁷. For transrepression experiments, wild-type PPAR- γ or PPAR- γ mutants were transfected at a 3:1 (PPAR- γ to reporter plasmid) ratio using Superfect reagent (Qiagen). For siRNA experiments, RAW264.7 cells were transfected with siRNAs (40 nM) using Superfect reagent for 48 h before activation with PPAR- γ ligands and LPS induction (6 h). In all transfections, cells were treated with 0.1 μ M rosiglitazone, stimulated with 1 μ g ml⁻¹ LPS and assayed for luciferase activity 6 h later. For mammalian two-hybrid assays, 200 ng of UAS-TK luciferase reporter and 100 ng each of VP16-PPAR- γ wild-type and Gal-DBD-NCOR constructs were transfected into RAW264.7 cells. VP16 refers to the transactivation domain of herpes simplex viral protein 16, and Gal-DBD refers to yeast Gal4 DNA-binding domain. Cells were cultured in 0.01 μ M trichostatin A (TSA) before ligand treatment for 16 h. Transfection experiments evaluated each experimental condition in triplicate and results are expressed as mean $\pm$ s.d. Statistical analysis was performed using Student's *t*-test, with *P* < 0.01 considered statistically significant.

Chromatin immunoprecipitation assays. ChIP assays were performed as previously described^{13,14}. For each experimental condition, 2–4 $\times$ 10⁶ primary macrophages or RAW 264.7 cells were used. Cells were pre-treated with 0.1 μ M rosiglitazone (1 h) and stimulated with 1 μ g ml⁻¹ LPS (1 h) before crosslinking for 10 min with 1% formaldehyde. Anti-PPAR- γ antibodies H-100 (Santa Cruz) and 39338 (Active Motif) were used in combination. Antibodies against TBL1 and TBLR1 have been described previously^{13,14}. Anti-HA protein A sepharose beads (Covance) were used for wild-type and mutant HA-tagged PPAR- γ proteins. Anti-HDAC3 and anti-p65 antibodies were from Santa Cruz. Anti-NCOR antibody was from Affinity Bioreagents. A 150-bp region of the iNOS promoter was amplified spanning the most proximal NF- κ B site to the start of transcription¹⁶. A 150-bp region of the mouse *Cd36* promoter was amplified spanning the PPRE sequence¹⁹.

RNA isolation, semiquantitative PCR and northern blot analysis. Total RNA (isolated by the Trizol method) was prepared from primary macrophages pretreated with 1 μ M rosiglitazone (2 h) before 1 μ g ml⁻¹ LPS stimulation (6 h). One microgram of total RNA was used for cDNA synthesis, and 2 μ l of cDNA was used for PCR using iNOS or inflammatory-gene-specific primers. For northern blot analysis, 10 μ g total RNA and an iNOS-specific probe were used.

Co-immunoprecipitations and western blotting. For co-immunoprecipitations, RAW264.7 cells or 293 cells were transfected with HA-PPAR- γ (wild type) and 2 $\times$ FLAG-PIAS1 (wild type) using Superfect reagent in 10-cm dishes. Whole-cell extracts were prepared using lysis buffer: 10 mM Tris-HCl pH 8, 420 mM NaCl, 1 mM EDTA and 0.5% NP-40 with protease inhibitor cocktail (Roche Biochem). Immunoprecipitates were washed four times with wash buffer containing 10 mM Tris-HCl pH 8, 100 mM NaCl, 1 mM EDTA, 0.5% NP-40 and 0.5% Triton X-100, boiled in 1 $\times$ sample loading buffer and separated by 10% SDS-PAGE. M2 anti-flag antibody (Sigma) was used at 1:1,000 dilution. Anti-HA antibody (Covance) was used at 1:1,000 dilution. Antibodies used for endogenous immunoprecipitation and western blot for PPAR- γ /PIAS1 interaction experiments were obtained from Active Motifs and Santa Cruz, respectively.

SUMOylation assays. For *in vivo* SUMOylation experiments, 250 μ g total protein extract was prepared from HeLa cells transfected with HA-tagged wild-type PPAR- γ or SUMO point mutants (K77R and K365R) and Myc-tagged SUMO-1. Cell lysates were immunoprecipitated and washed four times in lysis buffer containing 0.1% SDS, 0.5% deoxycholate, 0.5% TritonX-100, 1 mM EDTA, 20 mM Tris-HCl pH 7.8 and 150 mM NaCl. Immunoprecipitates were resolved by SDS-PAGE and immunoblotted using anti-HA or anti-Myc antibodies.

Received 10 June; accepted 30 June 2005.

Published online 28 August 2005.

1. Lehmann, J. M. *et al.* An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor γ (PPAR γ). *J. Biol. Chem.* **270**, 12953–12956 (1995).
2. Spiegelman, B. M. PPAR γ : adipogenic regulator and thiazolidinedione receptor. *Diabetes* **47**, 507–514 (1998).
3. Willson, T. M., Lambert, M. H. & Kliewer, S. A. Peroxisome proliferator-activated receptor γ and metabolic disease. *Annu. Rev. Biochem.* **70**, 341–367 (2001).
4. Haffner, S. M. *et al.* Effect of rosiglitazone treatment on nontraditional markers of cardiovascular disease in patients with type 2 diabetes mellitus. *Circulation* **106**, 679–684 (2002).
5. Li, A. *et al.* Peroxisome proliferator-activated receptor γ ligands inhibit development of atherosclerosis in LDL receptor-deficient mice. *J. Clin. Invest.* **106**, 523–531 (2000).
6. Chen, J. D. & Evans, R. M. A transcriptional co-repressor that interacts with nuclear hormone receptors. *Nature* **377**, 454–457 (1995).
7. Horlein, A. J. *et al.* Ligand-independent repression by the thyroid hormone receptor mediated by a nuclear receptor co-repressor. *Nature* **377**, 397–404 (1995).
8. Glass, C. K. & Rosenfeld, M. G. The coregulator exchange in transcriptional functions of nuclear receptors. *Genes Dev.* **14**, 121–141 (2000).
9. Guenther, M., Lane, W., Fischle, W., Verdin, E. & Lazar, M. A Core SMRT corepressor complex containing HDAC3 and TBL1, a WD40-repeat protein linked to deafness. *Genes Dev.* **14**, 1048–1057 (2000).
10. Li, J. *et al.* Both corepressor proteins SMRT and N-CoR exist in large protein complexes containing HDAC3. *EMBO J.* **19**, 4342–4350 (2000).
11. Zhang, J., Kalkum, M., Chait, B. T. & Roeder, R. G. The N-CoR-HDAC3 nuclear receptor corepressor complex inhibits the JNK pathway through the integral subunit GPS2. *Mol. Cell* **9**, 611–623 (2002).
12. Yoon, H. G. *et al.* Purification and functional characterization of the human N-CoR complex: the roles of HDAC3, TBL1 and TBLR1. *EMBO J.* **22**, 1336–1346 (2003).
13. Perissi, V., Aggarwal, A., Glass, C. K., Rose, D. W. & Rosenfeld, M. G. A corepressor/coactivator exchange complex required for transcriptional activation by nuclear receptors and other regulated transcription factors. *Cell* **116**, 511–526 (2004).
14. Ogawa, S. *et al.* A nuclear receptor corepressor transcriptional checkpoint controlling activator protein 1-dependent gene networks required for macrophage activation. *Proc. Natl Acad. Sci. USA* **101**, 14461–14466 (2004).
15. Hoberg, J. E., Yeung, F. & Mayo, M. W. SMRT derepression by the I κ B kinase α : A prerequisite to NF- κ B transcription and survival. *Mol. Cell* **16**, 245–255 (2004).
16. Lowenstein, C. J. *et al.* Macrophage nitric oxide synthase gene: two upstream regions mediate induction by interferon γ and lipopolysaccharide. *Proc. Natl Acad. Sci. USA* **90**, 9730–9734 (1993).
17. Ricote, M., Li, A. C., Willson, T. M., Kelly, C. J. & Glass, C. K. The peroxisome proliferator-activated receptor- γ is a negative regulator of macrophage activation. *Nature* **391**, 79–82 (1998).
18. Oberfield, J. L. *et al.* A peroxisome proliferator-activated receptor γ ligand inhibits adipocyte differentiation. *Proc. Natl Acad. Sci. USA* **96**, 6102–6106 (1999).
19. Tontonoz, P., Nagy, L., Alvarez, J. G. A., Thomazy, V. A. & Evans, R. M. PPAR γ promotes monocyte/macrophage differentiation and uptake of oxidized LDL. *Cell* **93**, 241–252 (1998).
20. Li, M., Pascual, G. & Glass, C. Peroxisome proliferator-activated receptor γ -dependent repression of the inducible nitric oxide synthase gene. *Mol. Cell. Biol.* **20**, 4699–4707 (2000).
21. Liu, B. *et al.* Inhibition of Stat1-mediated gene activation by PIAS1. *Proc. Natl Acad. Sci. USA* **95**, 10626–10631 (1998).
22. Jackson, P. K. A new RING for SUMO: wrestling transcriptional responses into nuclear bodies with PIAS family E3 SUMO ligases. *Genes Dev.* **15**, 3053–3058 (2001).
23. Tallec, L. P. *et al.* Protein inhibitor of activated signal transducer and activator of transcription 1 interacts with the N-terminal domain of mineralocorticoid receptor and represses its transcriptional activity: implication of small ubiquitin-related modifier 1 modification. *Mol. Endocrinol.* **17**, 2529–2542 (2003).
24. Janne, O. A. *et al.* Androgen-receptor-interacting nuclear proteins. *Biochem. Soc. Trans.* **28**, 401–405 (2000).
25. Tan, J. *et al.* Protein inhibitor of activated STAT-1 (signal transducer and activator of transcription-1) is a nuclear receptor coregulator expressed in human testis. *Mol. Endocrinol.* **14**, 14–26 (2000).
26. Kotaja, N., Karvonen, U., Janne, O. A. & Palvimäki, J. J. PIAS proteins modulate transcription factors by functioning as SUMO-1 ligases. *Mol. Cell. Biol.* **22**, 5222–5234 (2002).
27. Ling, Y. *et al.* Modification of de novo DNA methyltransferase 3a (Dnmt3a) by SUMO-1 modulates its interaction with histone deacetylases (HDACs) and its capacity to repress transcription. *Nucleic Acids Res.* **32**, 598–610 (2004).
28. Nishida, T. & Yasuda, H. PIAS1 and PIASx α function as SUMO-E3 ligases toward androgen receptor and repress androgen receptor-dependent transcription. *J. Biol. Chem.* **277**, 41311–41317 (2002).
29. Ohshima, T., Koga, H. & Shimotohno, K. Transcriptional activity of peroxisome proliferator-activated receptor γ is modulated by SUMO-1 modification. *J. Biol. Chem.* **279**, 29551–29557 (2004).
30. Muller, S., Hoege, C., Pyrowolakis, G. & Jentsch, S. SUMO, ubiquitin's mysterious cousin. *Nature Rev. Mol. Cell Biol.* **2**, 202–210 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Beas, W. Books and J. Goode for assistance with the yeast two-hybrid screen, and J. Collier for assistance with the mammalian two-hybrid assays. We thank F. Almazan for RAW264.7 cells. We thank M. Ricote for helpful discussions and M. Lambert for assistance with analysis of PPAR- γ crystal structures. We thank A. Howarth for assistance with preparation of the manuscript. These studies were supported by an American Heart Association Predoctoral Fellowship to G.P. and National Institutes of Health grants to D.W.R., M.G.R. and C.K.G. and support from the Stanford Donald W. Reynolds Cardiovascular Center to C.K.G. M.G.R. is supported by the Howard Hughes Medical Institute.

Author Contributions G.P., A.L.F., S.O., A.G., A.C.L. and D.W.R. performed experimental work and data analysis. T.M.W., M.G.R. and C.K.G. performed data analysis. C.K.G. was responsible for project planning.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to C.K.G. (cglass@ucsd.edu).

LETTERS

Structural basis of West Nile virus neutralization by a therapeutic antibody

Grant E. Nybakken¹, Theodore Oliphant², Syd Johnson⁵, Stephen Burke⁵, Michael S. Diamond^{1,2,4} & Daved H. Fremont^{1,3}

West Nile virus is a mosquito-borne flavivirus closely related to the human epidemic-causing dengue, yellow fever and Japanese encephalitis viruses¹. In establishing infection these icosahedral viruses undergo endosomal membrane fusion catalysed by envelope glycoprotein rearrangement of the putative receptor-binding domain III (DIII) and exposure of the hydrophobic fusion loop^{2–4}. Humoral immunity has an essential protective function early in the course of West Nile virus infection^{5,6}. Here, we investigate the mechanism of neutralization by the E16 monoclonal antibody that specifically binds DIII. Structurally, the E16 antibody Fab fragment engages 16 residues positioned on four loops of DIII, a consensus neutralizing epitope sequence conserved in West Nile virus and distinct in other flaviviruses. The E16 epitope protrudes from the surface of mature virions in three distinct environments⁷, and docking studies predict Fab binding will leave five-fold clustered epitopes exposed. We also show that E16 inhibits infection primarily at a step after viral attachment, potentially by blocking envelope glycoprotein conformational changes. Collectively, our results suggest that a vaccine strategy targeting the dominant DIII epitope may elicit safe and effective immune responses against flaviviral diseases.

To understand better the mechanism of antibody neutralization of West Nile virus we generated a large panel of envelope glycoprotein-specific monoclonal antibodies⁸. Domain mapping by yeast surface display revealed that ten out of twelve potentially neutralizing monoclonal antibodies selectively bind DIII. We also established that one of these monoclonal antibodies, E16, protects mice from lethal West Nile virus challenge even if administered therapeutically 5 days after infection. Here we have examined the structural basis for E16-mediated neutralization by determining the crystal structure of the Fab fragment in complex with West Nile virus DIII at 2.5 Å resolution (Supplementary Table S1). Our structure reveals that DIII adopts an immunoglobulin-like β -sandwich topology similar to that found in other flavivirus envelope glycoproteins, whereas the E16 Fab adopts a typical quaternary assembly (Fig. 1a). The binding interface has a high degree of shape complementarity ($S_c = 0.763$) (ref. 9) and occludes 1,550 Å² of surface area, with V_H (variable domain of heavy chain) accounting for 67% of the total antibody-combining site (Fig. 1b). E16 contacts DIII with 18 residues spread along all six of its CDR (complementarity determining region) loops in addition to three V_H framework residues (Supplementary Table S2). The interaction between E16 and DIII is dominated by hydrogen bonds, with 16 direct hydrogen bonds and numerous water-mediated networks at the interface of the complex.

E16 engages four discontinuous segments of DIII including the amino-terminal region (residues 302–309) and three strand-connecting loops: BC (330–333), DE (365–368) and FG (389–391). E16 contacts a total of 16 DIII residues, which together form a single

convex surface patch. Notably, yeast surface display epitope mapping¹⁰ of DIII identified four residues at the core of this binding site that are critical for E16 recognition (Fig. 1c). Non-conservative substitution at Ser 306, Lys 307, Thr 330 or Thr 332 disrupts E16 binding but not that of a non-neutralizing DIII-specific monoclonal antibody, E22. These four residues cluster at the centre of the E16–DIII interface (Fig. 1b). The decreased binding associated with mutation of the DIII residues Ser 306, Lys 307 and Thr 332 is most likely attributable to loss of hydrogen-bonding potential with E16, whereas Thr 330 appears to stabilize the DIII N-terminal strand conformation and provides numerous van der Waals contacts with the Fab (Fig. 1d, e). Collectively, our structural studies define the E16 epitope as a large surface patch on DIII created by four distinct secondary structure elements, and yeast mapping highlights the critical contributions of four central residues.

Comparison of available West Nile virus sequences reveals nearly complete conservation of the structurally defined E16 epitope (Fig. 2, and data not shown). Not surprisingly, E16 blocks infection of ten different lineage I and II West Nile virus strains⁸. Notably, nine other neutralizing monoclonal antibodies from our panel also lose the ability to recognize DIII after mutation of Ser 306, Lys 307, Thr 330 or Thr 332. Importantly, this epitope is also key in the humoral immune response of humans, as E16 Fabs effectively compete with West Nile virus convalescent antibodies for DIII binding⁸. Sequence analysis of other flaviviruses reveals a high degree of diversity in the four segments of the E16 epitope, with notable variation even between dengue serotypes (Fig. 2). Not surprisingly, E16 does not cross-neutralize dengue, Japanese or St Louis encephalitis viruses. Other groups have also identified flavivirus-specific neutralizing antibodies that localize to an analogous DIII-binding region^{11–13}. Thus, the coincident mapping of our and other neutralizing monoclonal antibodies suggests that the structural epitope has a dominant role in flavivirus neutralization.

To gain additional insight into the structural basis of E16-mediated neutralization, we docked our Fab–DIII complex onto the structures of the pre-fusion dengue envelope glycoprotein dimer¹⁴ and post-fusion trimer³ (Fig. 3a, b). The E16 epitopes are unencumbered in either configuration, although intact E16 antibody is unlikely to bivalently recognize these isolated oligomers due to extensive distal splaying of the Fab tails (Fig. 3a, b). DIII undergoes an $\sim 70^\circ$ rotation towards DII in the dimer to trimer transition^{2,3}, and E16 ligation per se could serve to hinder this conformational change. Moreover, E16 binds part of the linker that connects DIII to DI, and the N terminus of our truncated West Nile virus DIII fragment adopts a unique conformation that enables Tyr 302 to make contact with the E16 V_H domain (Fig. 3c). In envelope glycoprotein trimers, the flavivirus invariant Tyr 302 interacts with DI in a manner that would be disrupted if re-oriented as observed in the E16–DIII

¹Departments of Pathology & Immunology, ²Molecular Microbiology, ³Biochemistry and Molecular Biophysics, and ⁴Medicine, Washington University School of Medicine, St Louis, Missouri 63110, USA. ⁵MacroGenics, Rockville, Maryland 20850, USA.

complex. Thus, E16 binding could stabilize the mature state or alternatively restrict transition to the post-fusion conformation.

To understand better how E16 recognizes DIII in the context of the mature virus, we docked our structure onto the cryo-electron-microscopy-derived pseudo-atomic model of the intact West Nile virus virion^{7,16}. With three envelope glycoproteins in the asymmetric unit, there are three potential Fab-binding environments (Fig. 3d). Two binding modes are clearly allowed: one that closely circles the three-fold axis and a second disposed symmetrically about the icosahedral dyad that is permuted as an outer five-fold ring (Fig. 3e, f). However, the DIII epitopes are too tightly clustered at the true five-fold axis to permit E16 engagement without steric overlap with adjacent DIII residues. Thus, we postulate that at saturation no more than 120 Fabs can bind the 180 envelope glycoproteins in the mature virion, with exclusion of Fab binding to DIII around the inner five-fold ring (Fig. 3g). Additional exclusions may occur for intact antibody, although we note that E16 Fab alone neutralizes West Nile virus. This incomplete saturation is analogous to that observed in the decoration of papillomavirus by a neutralizing antibody, which blocks infection through a post-attachment mechanism¹⁷.

We next tested whether E16 could block cellular attachment of West Nile virus. Binding assays were performed with Vero cells, a cell

line permissive for West Nile virus infection. After a 4 °C incubation with West Nile virus in the presence of control (anti-SARS ORF7a), non-neutralizing (E22), or neutralizing monoclonal antibodies that map within (E16 or E24) or outside (E53 or E60) of DIII, cell-associated viral RNA was measured by fluorogenic polymerase chain reaction with reverse transcription (RT-PCR)¹⁸. Importantly, the non-binding and non-neutralizing monoclonal antibodies do not inhibit virus binding. In contrast, E53 and E60 block virus attachment by 8–9-fold ($P < 0.001$) whereas E16 and E24, which recognize the same dominant DIII epitope, only inhibit binding by 3.5-fold ($P = 0.003$) (Fig. 4a). The observation that E53 and E60 block virus binding more efficiently than E16 was not expected, as E53 and E60 are tenfold less potent in plaque reduction neutralization assays⁸.

Because E16 only partially blocks virus binding yet completely neutralizes infection, we tested whether E16 inhibits flavivirus infection by blocking a step after cellular attachment. Using a previously described assay^{19,20}, E16 or E53 was incubated with West Nile virus before, or after, mixing with a monolayer of Vero cells and infection was measured. Pre-binding of West Nile virus with either E16 or E53 significantly protects against infection (Fig. 4b). In contrast, E16 but not E53 significantly inhibits infection when added after virus binding. Because E16-mediated protection is not

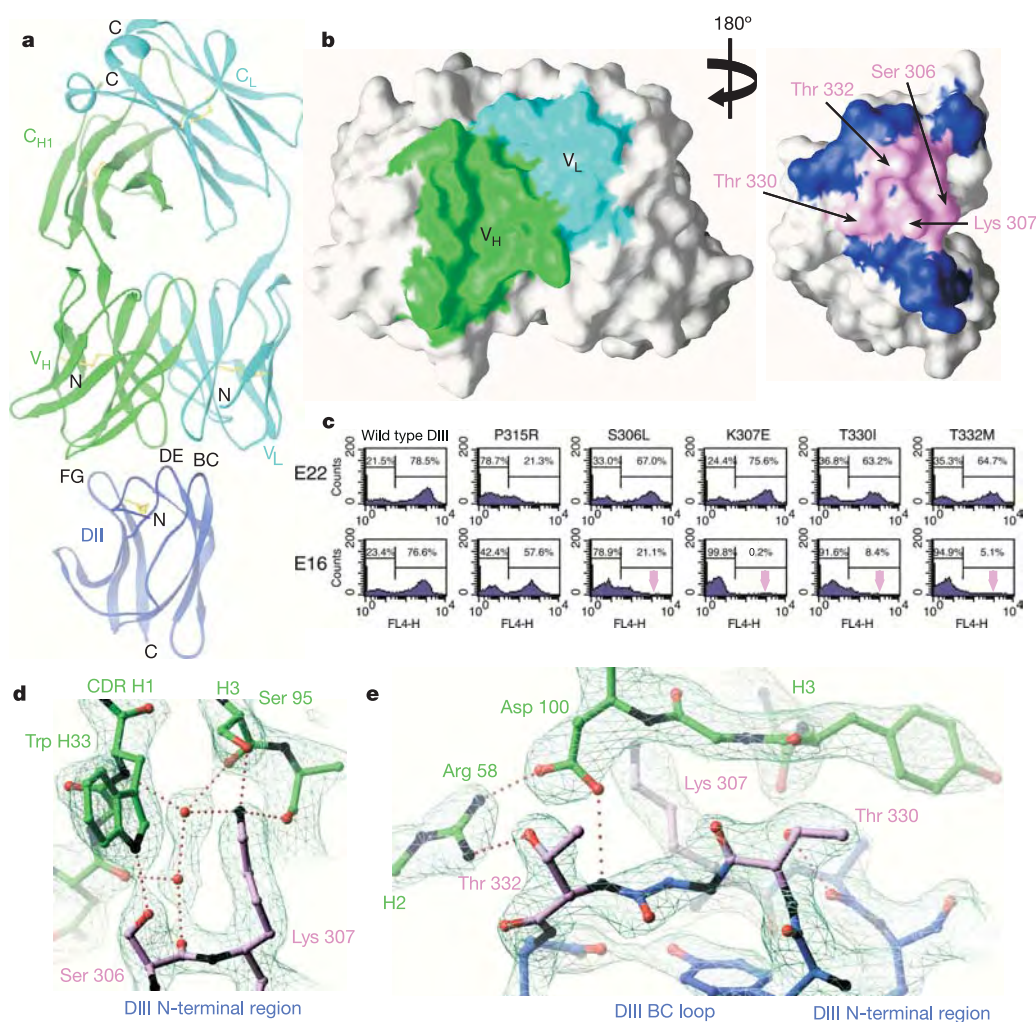


Figure 1 | Crystal structure of the E16 Fab in complex with DIII of West Nile virus envelope glycoprotein. **a**, Ribbon diagram of the complex, with DIII depicted in dark blue, the antibody heavy chain in green and the light chain in cyan. **b**, Molecular surface representation of the E16–DIII interface highlighting E16 V_H (green) and V_L (cyan) contact residues (left), as well as DIII contact residues (blue) and the residues defined by yeast surface display

(magenta) (right). **c**, Flow cytometry of yeast cells expressing wild-type or mutant versions of West Nile virus DIII. **d**, Detailed interactions of DIII residues Ser 306 and Lys 307 with E16, with interfacial waters (red) evident in the composite electron density omit map. **e**, Interactions of Thr 330 and Thr 332 at the E16 interface.

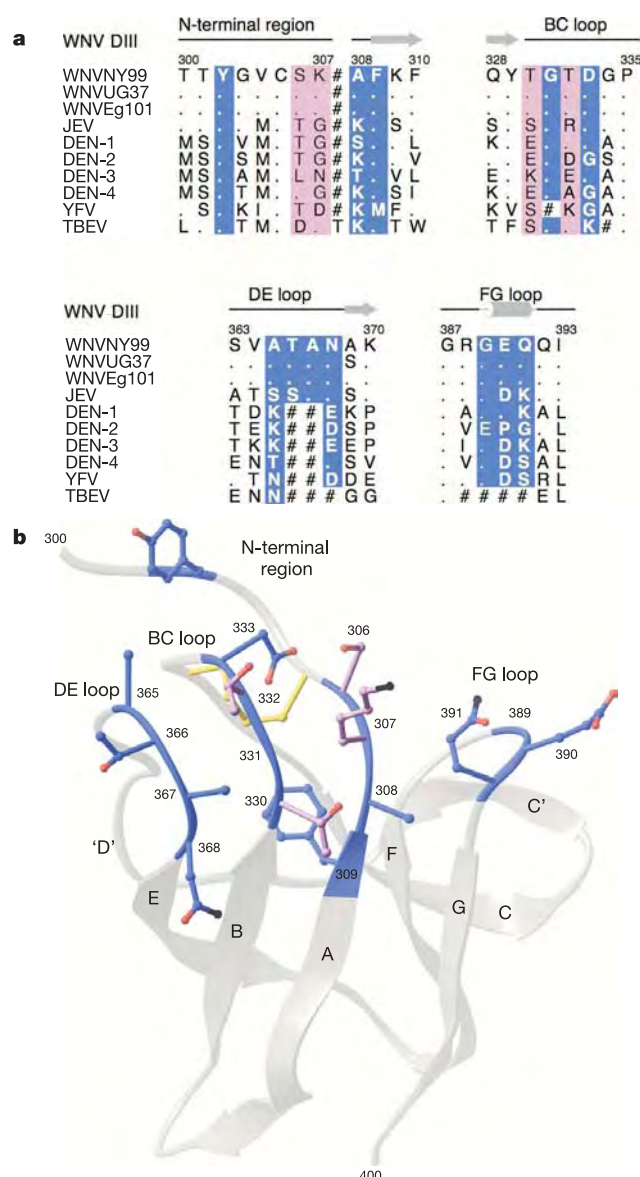


Figure 2 | The E16 binding site defines a West Nile virus dominant neutralizing epitope that is divergent in other flaviviruses. a, Sequence of the four segments of West Nile virus (WNV) DIII contacted by E16 aligned with the analogous residues of other flaviviruses. The DIII residues contacted by E16 are highlighted in blue and magenta, and deletions are indicated with a hash symbol. DEN, dengue virus; JEV, Japanese encephalitis virus; TBEV, tick-borne encephalitis virus; YFV, yellow fever virus. **b**, Structure of the West Nile virus dominant neutralizing epitope as defined by the E16–DIII complex.

appreciably affected by the time of addition, we surmise that it acts primarily after West Nile virus cellular attachment.

To define further the mechanism of West Nile virus neutralization, we evaluated whether E16 or other monoclonal antibodies enhance infection in macrophages. Antibody-dependent enhancement of infection occurs when antibody–virus complexes are preferentially internalized through Fc γ receptors on myeloid cells. Although the *in vivo* consequences remain uncertain, many monoclonal antibodies efficiently enhance flavivirus infection of Fc γ -receptor-bearing cells even when inhibitory in fibroblast neutralization assays²¹. We tested whether saturating concentrations of non-neutralizing (E5) or neutralizing (E16, E24 or E60) monoclonal antibodies enhance West Nile virus infection in macrophages. We found that whereas E5 and E60 augment infection 270- and 3,000-fold, respectively, E16 potently

inhibits macrophage infection at the same concentration. Notably, when E16 is combined with E5 or E60, it completely blocks enhancement as judged by reduction of virus yield (Fig. 4c) or viral RNA (data not shown). E24, which maps to the E16 dominant epitope, also blocks E5- and E60-dependent enhancement. Finally, the blockade of enhancement is not due to epitope competition as E16 and E24 do not cross-compete E5 or E60 for West Nile virus envelope glycoprotein binding (Fig. 4d and data not shown). Collectively, these virological experiments strongly suggest that E16 blocks West Nile virus infection primarily after cellular attachment.

We have shown that E16 binds a dominant neutralizing epitope on West Nile virus envelope glycoprotein defined by four distinct secondary structure elements that create a large surface patch on DIII, a region associated with pH-dependent conformational changes. Modelling studies suggest that E16 is excluded from five-fold clustered DIII epitopes on mature virions, potentially leaving them free to serve in receptor binding. Consistent with this, E16 inhibits West Nile virus infection primarily at a step after virus attachment. Moreover, our data suggest a potential advantage for interfering with post-attachment events, an emerging theme in other viral inhibition studies^{22–25}. Vaccines that selectively elicit antibodies specific to the critical DIII epitope may neutralize infection regardless of the mode of cellular entry, which may be particularly relevant given the increasing number of putative flavivirus entry receptors^{26–29}.

METHODS

Protein production and crystallization. West Nile virus DIII (residues 296–401) was expressed in bacteria with thrombin-cleavable histidine and BirA tags, and oxidatively re-folded from isolated inclusion bodies. Antibody–antigen complexes were formed by mixing papain-generated E16 Fab with thrombin-cleaved DIII, and purified by gel filtration chromatography. The E16–DIII complex crystallized at 15 mg ml^{−1} by hanging-drop vapour diffusion at 20 °C using 19% PEG 4000, 100 mM glycine and 100 mM HEPES (pH 8.5). Crystals were cryo-protected with precipitant solution containing 20% ethylene glycol, rapidly cooled in liquid nitrogen, and annealed by two 5-s blockages of the nitrogen stream.

Crystallographic structure determination. Data were collected at ALS beamline 4.2.2 (Lawrence Berkeley Laboratories) by the oscillation method at a wavelength of 1.55 Å at 100 K with a CCD detector. Data were processed, scaled and merged with d*TREK (<http://www.rigakumsc.com/protein/dtrek.html>). The crystals belong to space group *P*2₁2₁2₁, with unit cell dimensions of *a* = 52.4 Å, *b* = 83.3 Å and *c* = 110.6 Å, with one E16 Fab–DIII complex per asymmetric unit. Crystallographic phases were obtained by molecular replacement (<http://www.ccp4.ac.uk/>) using the coordinates of an IgG1 Fab' (Protein Data Bank (PDB) 2IGF) and the NMR structure of West Nile virus DIII (PDB 1S6N), which provided a correlation coefficient of 0.563 after rigid body refinement. An atomic model was iteratively built in O³⁰ and refined in CNS (<http://cns.csb.yale.edu/v1.1/>), and contains 532 amino acids (residues 300–400 of DIII, 1–212 of the E16 light chain and 1–228 of the E16 heavy chain) and 256 water molecules. The final 2.5 Å resolution model was refined to an *R*_{cryst} = 20.5% and *R*_{free} = 28.1% for all *F* > 0, with excellent geometry and Ramachandran angles (favoured, additional, generous and disallowed values of 87.5%, 11.9%, 0.4% and 0.2%, respectively).

Epitope mapping by yeast surface display. To define DIII residues critical for E16 recognition we expressed an error-prone PCR-derived library of West Nile virus envelope glycoprotein (residues 296–415) on the surface of yeast as Aga2 fusion proteins^{8,10}. Yeast were screened for selective loss of E16 binding relative to other DIII-specific monoclonal antibodies by multiple rounds of fluorescence-activated cell sorting. Isolated clones were recovered, sequenced and evaluated for binding to E16 and the non-neutralizing, West Nile virus DIII-specific monoclonal antibody E22.

West Nile virus binding to Vero cells. Individual purified monoclonal antibodies (50 µg ml^{−1} of anti-SARS ORF7a, E16, E22, E24, E53 or E60) or medium alone were incubated with 10³ plaque-forming units (p.f.u.) of West Nile virus for 1 h at 4 °C. These virus–antibody mixtures were then added to Vero cells in 12-well plates for 1 h on ice. Unbound virus was removed after six washes with PBS at 4 °C. Cells were lysed with guanidinium isothiocyanate, RNA was purified, and viral RNA was quantified by fluorogenic RT–PCR¹⁸.

Pre- and post-adsorption antibody inhibition assay. Increasing concentrations of E16 or E53 were added before or after West Nile virus (10² p.f.u.) binding (1 h

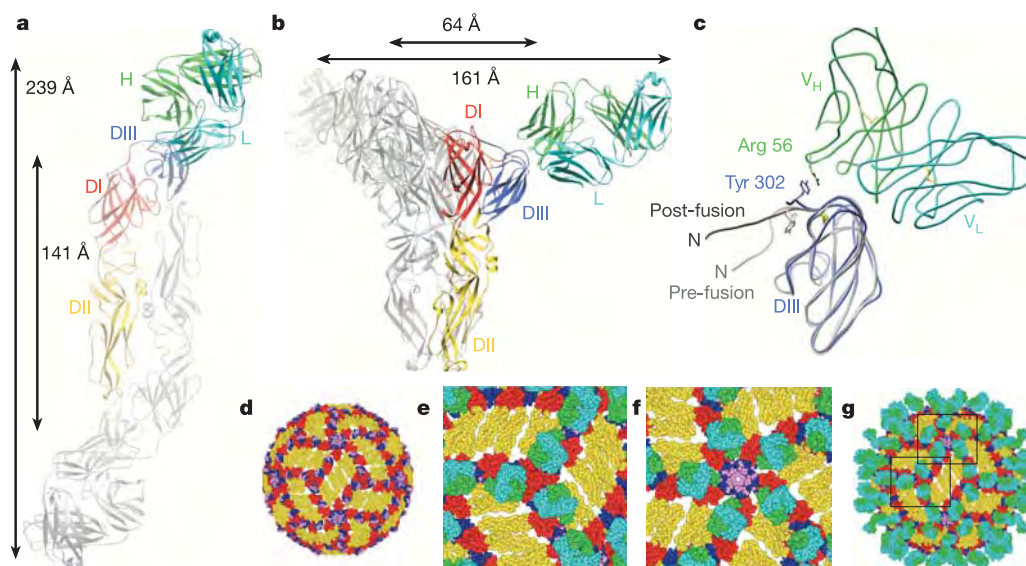


Figure 3 | E16 decoration of envelope glycoprotein assemblies. **a**, E16 docked onto the DEN-2 envelope glycoprotein dimer (1OAN) through DIII (blue). Super-positioning indicates that binding probably occurs without DI (red) or DII (yellow) contacts. The distances between DIII contact sites (141 Å) and the Fab termini (239 Å) preclude bivalent antibody engagement. **b**, E16 docked onto the post-fusion DEN-2 envelope glycoprotein trimer (1OK8) indicates accessibility of the binding site. **c**, DIII from the E16 complex (blue), the pre-fusion DEN-2 (1OKE, light grey) and post-fusion DEN-2 (1OK8, dark grey) reveal the conserved structure of the domains. The interaction of E16 with the flavivirus conserved Tyr 302 in the

N-terminal region of West Nile virus DIII is highlighted. **d**, The E16 structural epitope is mapped in magenta onto the cryo-electron microscopic reconstruction of the West Nile virus virion⁷ presented as 2.0-Å-radius α atoms. **e**, E16–DIII complexes docked around the icosahedral three-fold axis. **f**, DIII situated around the outer ring of the five-fold axis is permissive to E16 binding, but the inner ring appears to exclude E16 engagement. **g**, Saturation binding by E16 is predicted to entail the binding of 120 out of 180 West Nile virus epitopes with exclusion of binding to the inner-five-fold DIIIs (magenta). Panels **e** and **f** are magnified views of the boxed regions in **g**.

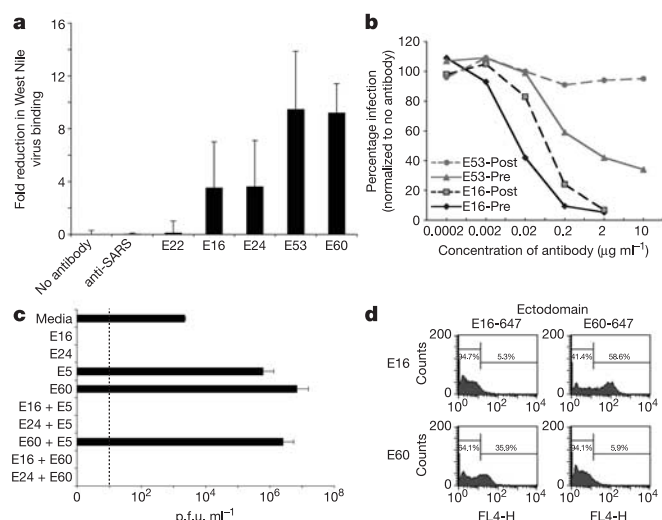


Figure 4 | Mechanism of E16-mediated neutralization of West Nile virus. **a**, Two DI/DII-specific neutralizing monoclonal antibodies (E53 and E60) block cellular attachment significantly more than the DIII-specific neutralizing antibodies (E16 and E24) or controls (no antibody, non-neutralizing monoclonal antibody E22 or anti-SARS ORF7a). Fold-reductions are reported, with standard deviations, as the average of four to seven independent experiments performed in triplicate. **b**, Dose-dependent blockade of West Nile virus infection by E16 and E53 in pre- and post-adsorption assays. The data are one of three representative experiments performed in duplicate. **c**, The DIII-specific monoclonal antibodies effectively inhibit West Nile virus infection of macrophages, whereas DI/DII-specific E5 and E60 monoclonal antibodies enhance infection. The data are one of three representative experiments performed in duplicate, with the dotted line representing the limit of sensitivity of the assay. Error bars represent the standard deviation. **d**, Pre-incubation with unlabelled monoclonal antibodies followed by addition of APC conjugates reveals that both E16 and E60 are self-competitive but not cross-competitive for envelope glycoprotein binding.

on ice) to Vero cells. In the post-adsorption assay, after washing away unbound virus, monoclonal antibody was allowed to bind for an additional hour. All cells were washed and an agarose overlay was added. Three days later, plaques were scored after fixation and staining with crystal violet.

Antibody-dependent enhancement assay. West Nile virus (5×10^2 p.f.u.) was pre-incubated with media, individual monoclonal antibodies ($50 \mu\text{g ml}^{-1}$ of E5, E16, E24 or E60) or combinations of monoclonal antibodies (E16 plus E5, E60 plus E5, E16 plus E60, E24 plus E5 or E24 plus E60) and then added to a monolayer (10^5) of J774.2 murine macrophages. After 6 h, cells were washed extensively with PBS to remove unbound virus and monoclonal antibody. After an additional 24 h, supernatants were harvested for a viral plaque assay on Vero cells.

Competitive binding of E16 and E60. Yeast expressing the empty vector pYD1 or the West Nile virus envelope glycoprotein ectodomain (residues 1–415) were incubated with $2.5 \mu\text{g}$ unlabelled E16 or E60 antibody for 1 h on ice. Unbound antibody was removed after three PBS washes containing 1 mg ml^{-1} BSA. E16 and E60 were conjugated using an Alexa Fluor 647 monoclonal antibody labelling kit (Molecular Probes). Conjugated E16 or E60 ($25 \mu\text{g ml}^{-1}$) was then added to the cells for 30 min at 4°C . Yeast were washed three times with PBS, fixed with 1% paraformaldehyde, and analysed using flow cytometry.

Received 9 April; accepted 24 June 2005.

- Granwehr, B. P. *et al.* West Nile virus: where are we now? *Lancet Infect. Dis.* **4**, 547–556 (2004).
- Bressanelli, S. *et al.* Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. *EMBO J.* **23**, 728–738 (2004).
- Modis, Y., Ogata, S., Clements, D. & Harrison, S. C. Structure of the dengue virus envelope protein after membrane fusion. *Nature* **427**, 313–319 (2004).
- Mukhopadhyay, S., Kuhn, R. J. & Rossmann, M. G. A structural perspective of the flavivirus life cycle. *Nature Rev. Microbiol.* **3**, 13–22 (2005).
- Diamond, M. S., Shrestha, B., Mehlich, E., Sitati, E. & Engle, M. Innate and adaptive immune responses determine protection against disseminated infection by West Nile encephalitis virus. *Viral Immunol.* **16**, 259–278 (2003).
- Wang, T. & Fikrig, E. Immunity to West Nile virus. *Curr. Opin. Immunol.* **16**, 519–523 (2004).
- Mukhopadhyay, S., Kim, B. S., Chipman, P. R., Rossmann, M. G. & Kuhn, R. J. Structure of West Nile virus. *Science* **302**, 248 (2003).
- Oliphant, T. *et al.* Development of a humanized monoclonal antibody with therapeutic potential against West Nile virus. *Nature Med.* **11**, 522–530 (2005).

9. Sundberg, E. J. & Mariuzza, R. A. Molecular recognition in antibody-antigen complexes. *Adv. Protein Chem.* **61**, 119–160 (2002).
10. Chao, G., Cochran, J. R. & Wittup, K. D. Fine epitope mapping of anti-epidermal growth factor receptor antibodies through random mutagenesis and yeast surface display. *J. Mol. Biol.* **342**, 539–550 (2004).
11. Wu, K. P. *et al.* Structural basis of a flavivirus recognized by its neutralizing antibody: Solution structure of the domain III of the Japanese Encephalitis virus envelope protein. *J. Biol. Chem.* **278**, 46007–46013 (2003).
12. Volk, D. E. *et al.* Solution structure and antibody binding studies of the envelope protein domain III from the New York strain of West Nile virus. *J. Biol. Chem.* **279**, 38755–38761 (2004).
13. Hiramatsu, K., Tadano, M., Men, R. & Lai, C. J. Mutational analysis of a neutralization epitope on the dengue type 2 virus (DEN2) envelope protein: monoclonal antibody resistant DEN2/DEN4 chimeras exhibit reduced mouse neurovirulence. *Virology* **224**, 437–445 (1996).
14. Modis, Y., Ogata, S., Clements, D. & Harrison, S. C. A ligand-binding pocket in the dengue virus envelope glycoprotein. *Proc. Natl Acad. Sci. USA* **100**, 6986–6991 (2003).
15. Saphire, E. O. *et al.* Crystal structure of a neutralizing human IgG against HIV-1: a template for vaccine design. *Science* **293**, 1155–1159 (2001).
16. Kuhn, R. J. *et al.* Structure of dengue virus: implications for flavivirus organization, maturation, and fusion. *Cell* **108**, 717–725 (2002).
17. Booy, F. P., Roden, R. B., Greenstone, H. L., Schiller, J. T. & Trus, B. L. Two antibodies that neutralize papillomavirus by different mechanisms show distinct binding patterns at 13 Å resolution. *J. Mol. Biol.* **281**, 95–106 (1998).
18. Diamond, M. S., Shrestha, B., Marri, A., Mahan, D. & Engle, M. B cells and antibody play critical roles in the immediate defense of disseminated infection by West Nile encephalitis virus. *J. Virol.* **77**, 2578–2586 (2003).
19. Crill, W. D. & Roehrig, J. T. Monoclonal antibodies that bind to domain III of dengue virus E glycoprotein are the most efficient blockers of virus adsorption to Vero cells. *J. Virol.* **75**, 7769–7773 (2001).
20. Hung, S. L. *et al.* Analysis of the steps involved in Dengue virus entry into host cells. *Virology* **257**, 156–167 (1999).
21. Halstead, S. B. & O'Rourke, E. J. Antibody-enhanced dengue virus infection in primate leukocytes. *Nature* **265**, 739–741 (1977).
22. Gollins, S. W. & Porterfield, J. S. A new mechanism for the neutralization of enveloped viruses by antiviral antibody. *Nature* **321**, 244–246 (1986).
23. Zolla-Pazner, S. Identifying epitopes of HIV-1 that induce protective antibodies. *Nature Rev. Immunol.* **4**, 199–210 (2004).
24. Skehel, J. J. & Wiley, D. C. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annu. Rev. Biochem.* **69**, 531–569 (2000).
25. Earp, L. J., Delos, S. E., Park, H. E. & White, J. M. The many mechanisms of viral membrane fusion proteins. *Curr. Top. Microbiol. Immunol.* **285**, 25–66 (2005).
26. Chen, Y. *et al.* Dengue virus infectivity depends on envelope protein binding to target cell heparan sulfate. *Nature Med.* **3**, 866–871 (1997).
27. Tassaneetrithep, B. *et al.* DC-SIGN (CD209) mediates dengue virus infection of human dendritic cells. *J. Exp. Med.* **197**, 823–829 (2003).
28. Halstead, S. B. Neutralization and antibody-dependent enhancement of dengue viruses. *Adv. Virus Res.* **60**, 421–467 (2003).
29. Tio, P. H., Jong, W. W. & Cardoso, M. J. Two dimensional VOPBA reveals laminin receptor (LAMR1) interaction with dengue virus serotypes 1, 2 and 3. *Virology* **338**, 25 (2005).
30. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for binding protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Alexander for yeast surface display expertise, O. Naidenko for expression plasmids, R. Kinney for the West Nile virus infectious clone, and R. Kuhn and M. Rossmann for coordinates of the West Nile virus mature virion. The authors also thank H. Virgin, T. Hansen and E. Unanue for critical comments on the manuscript. The work was supported by grants from the NIH and the Pediatric Dengue Vaccine Initiative (M.S.D. and D.H.F.). An NIH Medical Scientist Training Program Grant supported G.E.N.

Author Contributions G.E.N. carried out the protein biochemistry, crystallography and data analysis under the supervision of D.H.F. Epitope mapping and virology assays were performed by T.O. and M.S.D. E16 was sequenced by S.J. and Fabs were generated by S.B. The manuscript was written by G.E.N., M.S.D. and D.H.F.

Author Information E16 antibody sequences have been deposited under GenBank accession numbers DQ083997 (V_H) and DQ083998 (V_L). Coordinates for the E16 Fab–West Nile virus DIII complex have been deposited under accession code RCSB 1ZTX. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to D.H.F. (Fremont@wustl.edu).

A non-haem iron centre in the transcription factor NorR senses nitric oxide

Benoît D'Autr aux¹, Nicholas P. Tucker¹, Ray Dixon¹ & Stephen Spiro²

Nitric oxide (NO), synthesized in eukaryotes by the NO synthases, has multiple roles in signalling pathways and in protection against pathogens^{1,2}. Pathogenic microorganisms have apparently evolved defence mechanisms that counteract the effects of NO and related reactive nitrogen species. Regulatory proteins that sense NO mediate the primary response to NO and nitrosative stress^{3–9}. The only regulatory protein in enteric bacteria known to serve exclusively as an NO-responsive transcription factor is the enhancer binding protein NorR (refs 9–11). In *Escherichia coli*, NorR activates the transcription of the *norVW* genes encoding a flavorubredoxin (FIRd) and an associated flavoprotein, respectively, which together have NADH-dependent NO reductase activity^{10,12–14}. The NO-responsive activity of NorR raises important questions concerning the mechanism of NO sensing. Here we show that the regulatory domain of NorR contains a mononuclear non-haem iron centre, which reversibly binds NO. Binding of NO stimulates the ATPase activity of NorR, enabling the activation of transcription by RNA polymerase. The mechanism of NorR reveals an unprecedented biological role for a non-haem mononitrosyl-iron complex in NO sensing.

To gain insight into the nature of the active form of NorR, we sought evidence for the presence of an NO adduct in *E. coli* whole cells using electron paramagnetic resonance (EPR) spectroscopy. Whole cells exposed to NO under anaerobic conditions exhibited an EPR signal in the $g = 2$ region, which probably arises from complexes formed between NO and cellular and/or medium components (Fig. 1b, spectra 2 and 4). Cells expressing NorR and exposed to NO exhibited a new EPR signal in the $g = 4$ region, with two apparent g values at $g = 4.19$ and $g = 3.82$ (Fig. 1b, spectrum 4), which was absent from similarly treated cells either not expressing NorR (Fig. 1b, spectrum 2), or expressing NorR but not exposed to NO (Fig. 1b, spectrum 3). Thus, the formation of the $g = 4$ signal requires both the expression of NorR and treatment with NO. Similar EPR spectra have been observed after reaction of non-haem iron enzymes with NO (refs 15–19), suggesting that NorR possesses a non-haem iron centre.

Transcriptional activation by sigma 54 (σ^{54})-dependent activators is usually controlled by sensory modules¹¹, and in the case of NorR the amino-terminal GAF domain (named for cyclic GMP-specific and stimulated phosphodiesterases, *Anabaena* adenylate cyclases and *E. coli* FhlA) is the probable site of signal sensing (Fig. 1a). Deletion of the GAF domain (NorR Δ GAF) results in constitutive signal-independent transcriptional activation by NorR *in vivo* (Fig. 1e), implying that the function of the GAF domain is to sense the signal and inhibit the ATPase activity of the central AAA⁺ domain (Fig. 1a) when NorR is in its inactive state^{3,10}. Cells expressing only the GAF domain of NorR (GAF_{NorR}) and exposed to NO exhibited an EPR signal in the $g = 4$ region, identical to the signal observed with full-length NorR (Fig. 1c, spectrum 4). In contrast, cells expressing the GAF-truncated and constitutively active form of NorR, NorR Δ GAF,

did not exhibit the $g = 4$ EPR signal (Fig. 1d, spectrum 1), demonstrating that this signal does not arise from other proteins as a consequence of NorR-dependent activation of gene expression, and that the paramagnetic centre is located within the GAF domain of NorR.

NorR and GAF_{NorR} purified from overexpressing cells under aerobic conditions were both devoid of iron. However, when purified to homogeneity under anaerobic conditions (Supplementary Fig. 1a), NorR contained 0.3 Fe atoms per monomer. To increase the yield, we attempted the reconstitution of these preparations with ferrous iron in the absence of oxygen. After reconstitution and purification to remove free iron, GAF_{NorR} (Supplementary Fig. 1b) and full-length

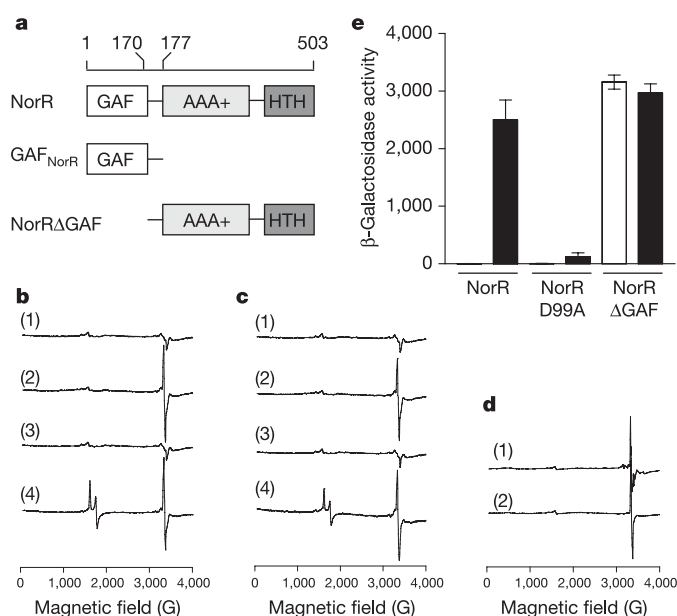


Figure 1 | Whole-cell EPR spectra and *in vivo* transcriptional activation by NorR. **a**, Schematic representation of the GAF, AAA⁺ and helix-turn-helix (HTH) domains of NorR and its derivatives, showing the relative amino acid positions of each domain. **b**, **c**, EPR spectra recorded from *E. coli* BL21 (DE3) cells expressing NorR (**b**) and GAF_{NorR} (**c**). In both instances, spectra were obtained from cultures treated as follows: spectra 1, no treatment; spectra 2, treatment with NO; spectra 3, induced for protein expression; spectra 4, induced for protein expression and then treated with NO. **d**, EPR spectra recorded from cultures expressing NorR Δ GAF (spectrum 1) and NorR D99A (spectrum 2), induced for protein expression and then treated with NO. **e**, Activation of *norV-lacZ* expression *in vivo* by plasmids carrying wild-type NorR, NorR D99A or NorR Δ GAF. Cultures were grown either in the absence (open bars) or presence (filled bars) of potassium nitrite as described in the Supplementary Methods. Error bars show standard deviation between repeat experiments.

¹John Innes Centre, Colney, Norwich NR4 7UH, UK. ²School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, Georgia 30332-0230, USA.

NorR contained 0.7 and 0.9 Fe atoms per monomer, respectively, suggesting that NorR coordinates a single mononuclear iron atom per monomer.

The iron centre in NorR seems to be in the reduced state because, in the absence of NO, EPR signals indicative of a ferric species were not detected either in whole cells (spectra 3 in Fig. 1b, c) or with purified reconstituted proteins (Supplementary Fig. 2). When exposed to NO, NorR-Fe(II) and GAF_{NorR}-Fe(II), unlike the apo-proteins, exhibited strong EPR signals characterized by the *g* values 4.19, 3.82 and 2.00, and *E/D* ≈ 0.03 (Fig. 2, spectra 1 and 2). The same EPR signal was obtained either with NO gas, or when using the pre-decomposed NO donor MAHMA NONOate (see Methods). The EPR signals are identical to those recorded from whole cells, indicating that the nitrosyl species observed *in vivo* can be entirely attributed to NO-modification of the Fe²⁺ centre in NorR. The set of *g* values is characteristic of five- and six-coordinated non-haem mononitrosyl {Fe(NO)}⁷ (*S* = 3/2) complexes^{15–19}, where the super-script refers to the number of valence electrons shared by the iron and NO, according to the Enemark and Feltham formalism²⁰. Furthermore, the identity of the EPR spectra obtained with full-length NorR and GAF_{NorR} (Fig. 2) indicates that all of the protein ligands to the nitrosyl-iron species are located in the GAF domain, and that the other domains of NorR do not influence the electronic properties of the mononitrosyl-iron centre. Titration of the nitrosyl-iron species by EPR revealed that 95% of the iron in both NorR and GAF_{NorR} is involved in the binding of NO. No significant differences were detected in the molecular mass of NorR (analysed by mass spectrometry under denaturing conditions) before and after treatment with NO (55,102 ± 2 and 55,103 ± 2 Da, respectively), demonstrating that NO does not form any additional covalent adduct with NorR (Supplementary Fig. 3).

To determine whether the iron centre in NorR is required for its activity, we substituted a candidate ligand to the iron, the strictly conserved aspartate residue at position 99 in the GAF domain of NorR, for alanine (NorR D99A). This mutation prevented the formation of the *g* = 4 EPR signal, both *in vivo* (Fig. 1d, spectrum 2) and *in vitro* (Fig. 2, spectrum 3). In contrast to wild-type NorR, the anaerobically purified NorR D99A mutant protein did not contain iron. Moreover, unlike wild-type NorR, the NorR D99A mutant was unable to activate expression from a *norVW-lacZ* reporter construct

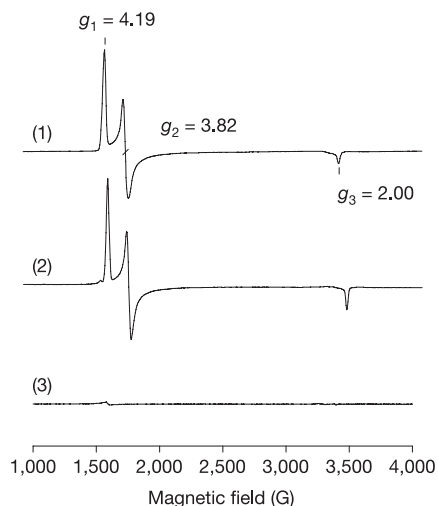


Figure 2 | EPR spectra of NO-treated purified proteins. Spectrum 1, purified NorR-Fe(II); spectrum 2, GAF_{NorR}-Fe(II); and, spectrum 3, NorR D99A. Protein concentrations were 40 μM for NorR-Fe(II), 90 μM for GAF_{NorR}-Fe(II) and 100 μM for NorR D99A. Identical spectra were obtained when proteins were treated with two equivalents of NO from a predecomposed MAHMA NONOate solution, or with NO gas.

in vivo (Fig. 1e), demonstrating that the loss of Fe²⁺ binding correlates with the absence of NorR activity.

The central catalytic AAA⁺ domain of enhancer binding proteins is required to couple nucleotide hydrolysis to the formation of open promoter complexes by σ⁵⁴-RNA polymerase²¹. Because the transcriptional activation mediated by this domain of NorR is apparently inhibited by the GAF domain in the absence of the signal, we anticipated that the coordination of NO to the Fe(II) centre might activate NorR. As a positive control for activity assays, we used the truncated form of NorR (NorRΔGAF), which lacks the GAF domain and gives rise to signal-independent transcriptional activation *in vivo* (Fig. 1a, e)^{3,10}. Isomerization of the σ⁵⁴-RNA polymerase holoenzyme was detected as a heparin-resistant super-shifted nucleoprotein complex on non-denaturing gels (Fig. 3a, lanes 8 and 9). Formation of the super-shifted species required not only NorRΔGAF, σ⁵⁴-RNA polymerase, integration host factor (IHF) and ATP, but also an initiating nucleotide, with CTP being more effective than UTP (Fig. 3a). NorR apoprotein prepared under aerobic conditions

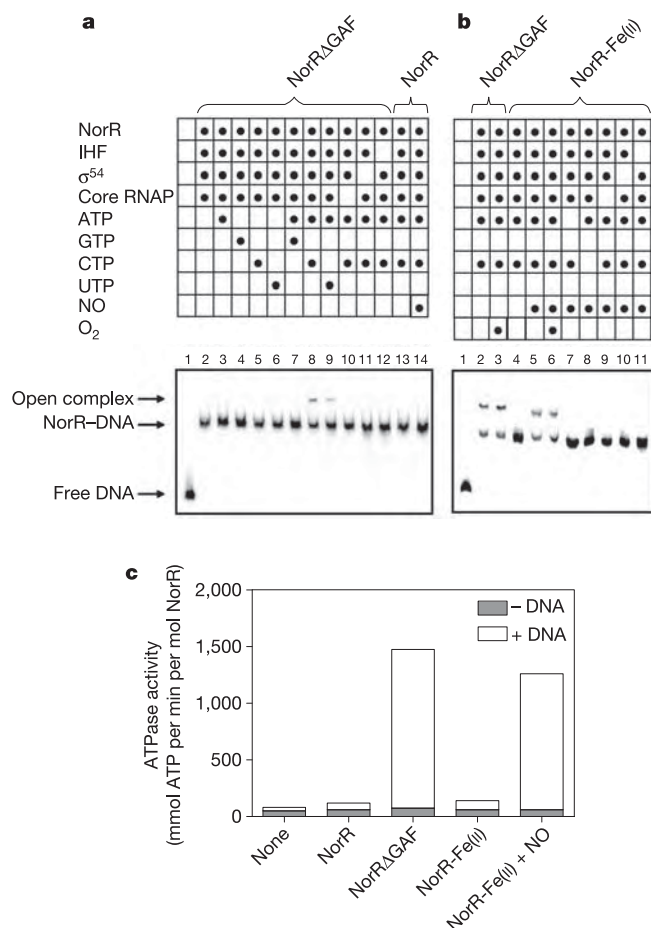


Figure 3 | Formation of open promoter complexes and the ATPase activity of NorR. **a**, Open promoter complex assays with NorRΔGAF (lanes 2–12) and NorR apoprotein (lanes 13 and 14). Reactions were performed under anaerobic conditions with the components indicated by the bullet points above each lane, in addition to DNA and buffer. RNAP, RNA polymerase. **b**, Open promoter complex formation by NorRΔGAF (lanes 2 and 3) and NorR-Fe(II) (lanes 4–11). Reactions were performed under anaerobic conditions except in lanes 3 and 6, which were incubated in air. **c**, ATPase activities of NorR derivatives. Rates of ATP hydrolysis were monitored at 340 nm for 20 min at 37 °C (filled bars). Then, 5 nM of the *norR-norVW* DNA fragment was added and the rates were monitored for an additional 20 min at 37 °C (open bars). ATPase activities are expressed as specific activity relative to protein concentration and the bars represent the sum of the values obtained without and with DNA.

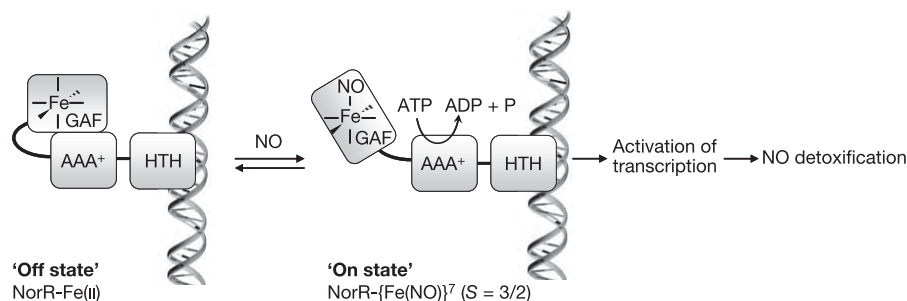


Figure 4 | Schematic of the proposed mechanism for transcriptional activation by NorR. NorR is represented in the 'off state' as a NorR-Fe(II)-DNA complex, and is activated by NO to an 'on state' with ATPase activity. NorR is shown with a tripartite domain architecture comprising the carboxy-terminal HTH DNA binding domain, the central catalytic AAA⁺ domain and the N-terminal GAF domain containing a five- or six-coordinate

mononuclear non-haem ferrous iron ion (here shown as a six-coordinated iron centre for clarity). The binding of NO leads to a {Fe(NO)}⁷ (*S* = 3/2) mononitrosyl complex where the interaction between the GAF domain and the AAA⁺ domain is released, allowing ATP hydrolysis that is further coupled to transcriptional activation. For clarity, NorR is represented as a monomer.

(and thus devoid of iron) was unable to drive the formation of open promoter complexes either in the presence or absence of NO (Fig. 3a, lanes 13 and 14). In contrast, a super-shifted species with mobility similar to that of the open complex formed in the presence of NorRΔGAF was seen in reactions containing NorR-Fe(II) treated with NO (Fig. 3b, lanes 2 and 3 versus lanes 5 and 6). These data demonstrate that both the iron centre and NO are required for open complex formation and that no other cofactors are involved in NorR-dependent activation. Open complexes were formed in reactions performed in air (Fig. 3b, lane 6), and EPR spectra of NO-treated ferrous NorR were not modified after exposure to air. Therefore, NO-activated NorR-Fe(II) is apparently not sensitive to oxygen, which is consistent with the presence of NorR activity in aerobic cultures^{9,14}.

The requirement for ATP in the open complex assays probably reflects the need for nucleotide hydrolysis by the AAA⁺ domain of NorR in remodelling the σ^{54} -RNA polymerase-promoter complex to activate transcription²². We therefore assayed the ATPase activity of NorR to obtain further evidence that the protein is activated by the formation of a nitrosyl-iron complex. All forms of NorR, including the constitutively active NorRΔGAF, NorR-Fe(II) and NO-treated NorR-Fe(II), had similar basal levels of activity. However, the ATPase activities of NorRΔGAF and NO-treated NorR-Fe(II) markedly increased in the presence of DNA containing NorR binding sites (Fig. 3c). Stimulation of ATPase activity on binding to specific enhancers has also been demonstrated for some other σ^{54} -dependent activators²³. The enhancer dependency of the ATPase activity potentially increases the specificity of NO signalling by NorR towards the regulation of specific target promoters. Clearly, the DNA-dependent ATPase activity of full-length NorR requires activation of the iron-containing species by NO, which may account for the signal-dependent isomerization of σ^{54} -RNA polymerase that we observe in the open complex assays.

Using an NO electrode, we obtained a dissociation constant, $K_d = 50 \pm 10$ nM at pH 8.5 and 30 °C (Supplementary Fig. 4), which indicates that NorR is sufficiently sensitive to respond to the micromolar levels of NO released by macrophages during inflammation²⁴, and the nanomolar levels made endogenously by nitrite reduction (Supplementary Methods). Determination of approximate rate constants for NO binding, with myoglobin as a competitor ($k_{\text{off}} > 1 \text{ s}^{-1}$ and $k_{\text{on}} > 2 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$), suggests that NorR is a rapid and efficient NO sensor.

We have identified a mononuclear non-haem iron centre in the GAF domain of NorR, which can be modified by a single NO molecule to generate a mononitrosyl {Fe(NO)}⁷ (*S* = 3/2) species. We propose that formation of this species triggers a conformational change in the GAF domain that relieves intramolecular repression of the AAA⁺ domain (Fig. 4), allowing enhancer-stimulated ATPase activity, open complex formation and transcription initiation by

σ^{54} -RNA polymerase. This mechanism enables direct transduction of the NO signal by NorR to the activation of NO detoxification processes. Diverse signal transduction pathways involving GAF domains have been described²⁵, but NorR provides the first example of a GAF domain coordinating a transition metal to sense and transduce a signal. EPR spectroscopy showed that the nitrosyl complex in the GAF domain of NorR belongs to the {Fe(NO)}⁷ (*S* = 3/2) family. To our knowledge, activation of NorR is the first example of a biological process involving a {Fe(NO)}⁷ (*S* = 3/2) complex. Previously reported NO signalling mechanisms involve either dinitrosyl-iron {Fe(NO)₂}⁹ complexes^{5,6,8,26}, or (as in the case of soluble guanylate cyclase) a haem-containing mononitrosyl complex of the {Fe(NO)}⁷ (*S* = 1/2) type²⁷. In contrast to most dinitrosyl-iron complexes²⁶, NO binding is reversible in mononitrosyl systems. Mononitrosyl low-spin (*S* = 1/2) haem systems and mononitrosyl high-spin (*S* = 3/2) non-haem systems are clearly distinct owing to their electronic and spatial structures^{18,19,28,29}, which may influence NO binding properties, ligand selectivity and mechanisms of NO signal transduction. The system we describe here has no precedent in NO signalling pathways and may provide a paradigm for NO coordination in NO receptors containing non-haem iron.

METHODS

Bacterial strains and plasmids. The pNorR2 plasmid containing the *norR* gene of *E. coli* in the pET21a vector has been described previously¹². Two derivatives of pNorR2 containing genes that express deleted versions of NorR were constructed. One is a deletion of the N-terminal domain from amino acids 1 to 170, and expresses the protein designated NorRΔGAF. The second expresses the NorR GAF domain only (designated GAF_{NorR}), from residues 1 to 177. For open complex and ATPase activity assays, the *norR-norVW* intergenic region was amplified and cloned into the *Sma*I site of pUC19 to yield the pNPTprom construct. Restriction digestion of pNPTprom with *Eco*RI and *Bam*HI generates a 362 base pair (bp) fragment spanning the entire *norR-norVW* intergenic region and includes the first 86 bp and 62 bp of the *norR* and *norV* coding sequences, respectively.

Nitric oxide treatment. Two different sources of NO were used to prepare NO solutions: the NO donor, MAHMA NONOate ((*Z*)-1-[*N*-Methyl-*N*-(6-(*N*-methylammoniohexyl)amino)]diazene-1-ium-1,2-diolate), and NO gas. Stock solutions of MAHMA NONOate were prepared in 10 mM NaOH. NO solutions were prepared by decomposition of a solution of MAHMA NONOate in buffer I (100 mM Tris-HCl pH 8.5, 100 mM NaCl, 5% glycerol) at 30 °C for 20 min ($t_{1/2} = 3$ min, pH 7.5, 30 °C) under anaerobic conditions. NO gas was purified by bubbling through 1 M KOH. Saturated solutions (2 mM) were prepared by bubbling pure NO gas into buffer I in anaerobic conditions.

Electron paramagnetic resonance experiments. For whole-cell experiments, freshly transformed bacteria were grown anaerobically at 30 °C in 25 ml of Luria-Bertani medium containing 1% (w/v) glucose and 100 μg ml⁻¹ carbenicillin. Expression of NorR and GAF_{NorR} was induced at an absorbance ($A_{600 \text{ nm}}$) of 0.6 with 50 μM isopropyl-β-D-thiogalactoside. After 3 h, 50 μl of 100 mM MAHMA NONOate was added and incubation continued for 15 min ($t_{1/2} = 3$ min,

pH 7.5, 30 °C). The cells were collected at 1,800g for 10 min, and the pellets were resuspended in 0.5 ml of 100 mM Tris-HCl (pH 7.5), 25 mM NaCl, 10% (v/v) glycerol. Cell suspensions were transferred to an EPR tube and immediately frozen. For EPR spectroscopy of purified proteins, samples were prepared by reaction of NorR-Fe(II) and GAF_{NorR}-Fe(II) with two equivalents of NO from predecomposed MAHMA NONOate or NO gas in anaerobic conditions. To observe the $g = 2$ contribution, the solution was incubated for a further 2 min with the cap open, to allow free NO to equilibrate with the gas phase, and then the solution was frozen. After this incubation, free NO was not observed in solution. X-band EPR spectra were recorded on a Bruker ER 200 D-SRC spectrometer under the following conditions: frequency, 9.477 GHz (for whole cells) or 9.440 GHz (for pure protein); power, 2 mW; modulation amplitude, 10 G; modulation frequency, 100 kHz; temperature 8 K. Spin concentrations were measured by integration of the EPR absorption spectra. The resulting areas were compared to the signal from aqueous Cu(H₂O)₆ (1 mM) recorded with identical instrument settings. EPR simulations were obtained with WinEPR Simfonia software (Bruker).

Protein purification. All proteins were purified to homogeneity (Supplementary Methods and Supplementary Fig. 1). Protein concentrations were determined by quantitative amino acid analysis (Alta Biosciences). According to these quantitative analyses, we have determined the absorption coefficients, at a wavelength of 280 nm, to be 42,000 M⁻¹ cm⁻¹, 6,500 M⁻¹ cm⁻¹ and 35,000 M⁻¹ cm⁻¹ for NorR, GAF_{NorR} and NorRΔGAF, respectively. Total iron concentrations were determined by inductively coupled plasma mass spectrometry (Harwell Scientific).

Open complex assays. Open complex formation was assayed as described previously³⁰ under anaerobic conditions using a modified buffer system to avoid iron chelation (Supplementary Methods).

ATPase activity assays. ATPase activity was measured using an assay in which the production of ADP is coupled to the oxidation of NADH by lactate dehydrogenase and pyruvate kinase. The oxidation of NADH was monitored by A_{340 nm} at 37 °C for 20 min. All reaction mixtures contained ATP (30 mM), phosphoenolpyruvate (1 mM), NADH (0.3 mM), pyruvate kinase (7 U), lactate dehydrogenase (23 U) in 50 mM Tris-HCl (pH 8.0), 100 mM KCl, 2 mM MgCl₂ and 300 nM of either NorR, NorRΔGAF or NorR-Fe(II). Where indicated, NorR-Fe(II) was pretreated with 20 μM NO. The DNA template added was the same *norR-norVW* promoter fragment (5 nM) as used for the open complex assays and is described in the Supplementary Methods.

Other protocols. Detailed procedures used for site-directed mutagenesis and assays of NorR activity *in vivo*, mass spectrometry, determination of the NO dissociation constant and kinetic parameters, open complex assays, protein purifications and details of chemical and biochemicals used are provided in the Supplementary Methods.

Received 18 March; accepted 21 June 2005.

- Griffith, O. W. & Stuehr, D. J. Nitric oxide synthases: properties and catalytic mechanism. *Annu. Rev. Physiol.* **57**, 707–736 (1995).
- MacMicking, J., Xie, Q. W. & Nathan, C. Nitric oxide and macrophage function. *Annu. Rev. Immunol.* **15**, 323–350 (1997).
- Pohlmann, A., Cramm, R., Schmelz, K. & Friedrich, B. A novel NO-responding regulator controls the reduction of nitric oxide in *Ralstonia eutropha*. *Mol. Microbiol.* **38**, 626–638 (2000).
- Fang, F. C. Antimicrobial reactive oxygen and nitrogen species: Concepts and controversies. *Nature Rev. Microbiol.* **2**, 820–832 (2004).
- D'Autréaux, B., Touati, D., Bersch, B., Latour, J. M. & Michaud-Soret, I. Direct inhibition by nitric oxide of the transcriptional ferric uptake regulation protein via nitrosylation of the iron. *Proc. Natl Acad. Sci. USA* **99**, 16619–16624 (2002).
- Ding, H. G. & Dimple, B. Direct nitric oxide signal transduction via nitrosylation of iron-sulfur centers in the SoxR transcription activator. *Proc. Natl Acad. Sci. USA* **97**, 5146–5150 (2000).
- Hausladen, A., Privalle, C. T., Keng, T., DeAngelo, J. & Stamler, J. S. Nitrosative stress: activation of the transcription factor OxyR. *Cell* **86**, 719–729 (1996).
- Cruz-Ramos, H. *et al.* NO sensing by FNR: regulation of the *Escherichia coli* NO-detoxifying flavohaemoglobin, Hmp. *EMBO J.* **21**, 3235–3244 (2002).
- Mukhopadhyay, P., Zheng, M., Bedzyk, L. A., LaRossa, R. A. & Storz, G. Prominent roles of the NorR and Fur regulators in the *Escherichia coli* transcriptional response to reactive nitrogen species. *Proc. Natl Acad. Sci. USA* **101**, 745–750 (2004).
- Gardner, A. M., Gessner, C. R. & Gardner, P. R. Regulation of the nitric oxide reduction operon (*norRVW*) in *Escherichia coli*. Role of NorR and σ^{54} in the nitric oxide stress response. *J. Biol. Chem.* **278**, 10081–10086 (2003).
- Studholme, D. J. & Dixon, R. Domain architectures of σ^{54} -dependent transcriptional activators. *J. Bacteriol.* **185**, 1757–1767 (2003).
- Tucker, N. P., D'Autréaux, B., Studholme, D. J., Spiro, S. & Dixon, R. DNA binding activity of the *Escherichia coli* nitric oxide sensor NorR suggests a conserved target sequence in diverse Proteobacteria. *J. Bacteriol.* **186**, 6656–6660 (2004).
- Gomes, C. M. *et al.* A novel type of nitric-oxide reductase. *Escherichia coli* flavorubredoxin. *J. Biol. Chem.* **277**, 25273–25276 (2002).
- Hutchings, M. I., Mandhana, N. & Spiro, S. The NorR protein of *Escherichia coli* activates expression of the flavorubredoxin gene *norV* in response to reactive nitrogen species. *J. Bacteriol.* **184**, 4640–4643 (2002).
- Arciero, D. M., Lipscomb, J. D., Huynh, B. H., Kent, T. A. & Munck, E. EPR and Mössbauer studies of protocatechuate 4,5-dioxygenase. Characterization of a new Fe²⁺ environment. *J. Biol. Chem.* **258**, 14981–14991 (1983).
- Clay, M. D., Cosper, C. A., Jenney, F. E. Jr, Adams, M. W. & Johnson, M. K. Nitric oxide binding at the mononuclear active site of reduced *Pyrococcus furiosus* superoxide reductase. *Proc. Natl Acad. Sci. USA* **100**, 3796–3801 (2003).
- Ray, M. *et al.* Structure and magnetic properties of trigonal bipyramidal iron nitrosyl complexes. *Inorg. Chem.* **38**, 3110–3115 (1999).
- Brown, C. A. *et al.* Spectroscopic and theoretical description of the electronic structure of $S = 3/2$ iron-nitrosyl complexes and their relation to O₂ activation by non-heme iron enzyme active-sites. *J. Am. Chem. Soc.* **117**, 715–732 (1995).
- Hauser, C., Glaser, T., Bill, E., Weyhermüller, T. & Wieghardt, K. The electronic structures of an isostructural series of octahedral nitrosyliron complexes {Fe-NO}^{6,7,8} elucidated by Mössbauer spectroscopy. *J. Am. Chem. Soc.* **122**, 4352–4365 (2000).
- Enemark, J. H. & Feltham, R. D. Principles of structure, bonding, and reactivity for metal nitrosyl complexes. *Coord. Chem. Rev.* **13**, 339–406 (1974).
- Cannon, W. V., Gallegos, M. T. & Buck, M. Isomerization of a binary sigma-promoter DNA complex by transcription activators. *Nature Struct. Biol.* **7**, 594–601 (2000).
- Zhang, X. *et al.* Mechanochemical ATPases and transcriptional activation. *Mol. Microbiol.* **45**, 895–903 (2002).
- Austin, S. & Dixon, R. The prokaryotic enhancer binding protein NTRC has an ATPase activity which is phosphorylation and DNA dependent. *EMBO J.* **11**, 2219–2228 (1992).
- Roy, S., Sharma, S., Sharma, M., Aggarwal, R. & Bose, M. Induction of nitric oxide release from the human alveolar epithelial cell line A549: an *in vitro* correlate of innate immune response to *Mycobacterium tuberculosis*. *Immunology* **112**, 471–480 (2004).
- Aravind, L. & Ponting, C. P. The GAF domain: an evolutionary link between diverse phototransducing proteins. *Trends Biochem. Sci.* **22**, 458–459 (1997).
- D'Autréaux, B. *et al.* Spectroscopic description of the two nitrosyl-iron complexes responsible for Fur inhibition by nitric oxide. *J. Am. Chem. Soc.* **126**, 6005–6016 (2004).
- Zhao, Y., Brandish, P. E., Ballou, D. P. & Marletta, M. A. A molecular basis for nitric oxide sensing by soluble guanylate cyclase. *Proc. Natl Acad. Sci. USA* **96**, 14753–14758 (1999).
- Li, M. *et al.* Tuning the electronic structure of octahedral iron complexes [FeL(X)] (L = 1-alkyl-4,7-bis(4-*tert*-butyl-2-mercaptobenzyl)-1,4,7-triazacyclononane, X = Cl, CH₃O, CN, NO). The $S = 1/2 \rightleftharpoons S = 3/2$ spin equilibrium of [FeL⁺(NO)]. *Inorg. Chem.* **41**, 3444–3456 (2002).
- Seres, R. G. *et al.* Structural, spectroscopic, and computational study of an octahedral, non-heme [Fe-NO]^{6–8} series: [Fe(NO)(cyclam-ac)]^{2+/+0}. *J. Am. Chem. Soc.* **126**, 5138–5153 (2004).
- Little, R., Reyes-Ramirez, F., Zhang, Y., van Heeswijk, W. C. & Dixon, R. Signal transduction to the *Azotobacter vinelandii* NIFL-NIFA regulatory system is influenced directly by interaction with 2-oxoglutarate and the PII regulatory protein. *EMBO J.* **19**, 6041–6050 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was funded by a grant from the BBSRC to R.D. and S.S. We are grateful to R. Little, G. Savers, M. Cheesman, I. Martinez-Argudo, P. Johnson and S. Fairhurst for their assistance and comments at various stages of this project, to M. Naldrett and A. Bottrill for their help with mass spectrometry, and to M. Buttner for comments on the manuscript.

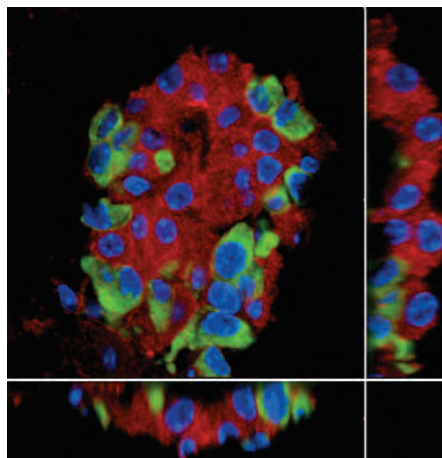
Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.D. (ray.dixon@bbsrc.ac.uk) or S.S. (stephen.spiro@biology.gatech.edu).

The big picture

Over the past ten years, microscopy has been transformed from slice, stain and fix, to the capacity to view living cells and even whole organisms in real time. **Lisa Melton** looks at what's on offer.

NIKON When it was introduced in the late 1980s, confocal laser-scanning microscopy opened up a new high-resolution world to biologists. But researchers are becoming more demanding of their microscopes. They expect not just to see a clearer image, but to monitor dynamic protein networks within cells, map the kinetics of intracellular organelles and track calcium signalling. Microscope developers have responded by producing confocal microscope systems that have dramatically improved scanning speeds, greater resolution and the capacity to see detail in live cells labelled with multiple colours.

Researchers traditionally struggled to obtain sharp images from multi-stained specimens, the main obstacle being spectral emission overlap from different fluorescent probes. Nikon's recently launched Digital Eclipse C1 Spectral Imaging confocal system collects high-resolution data from the 400–750 nanometres range with a single pass of the laser. A mathematical process unmixes closely overlapping spectral data to produce clean images with no cross-talk, even for notoriously difficult reds. "We use chromatic aberration-free objectives, so everything towards either end of the spectra focuses



Pancreatic cells imaged in the Digital Eclipse.

at the same point," says Chay Keogh, marketing manager for Nikon, UK in Kingston upon Thames. "By eliminating the need for multiple scans, specimen damage is kept to a minimum."

A rival system from Olympus, the Fluoview FV1000, is a confocal laser-scanning microscope with two independent, synchronized laser scanners in a single instrument. While

one laser provides high-resolution confocal images, the second scanner, the SIM scanner, simultaneously stimulates the sample. This makes the FV1000 an ideal choice for live-cell applications such as fluorescence recovery after photobleaching (FRAP), fluorescence loss in photobleaching (FLIP), uncaging, or photoactivation and photoconversion. "It offers, for the first time, the opportunity to study the kinetics of rapid cellular reactions after laser stimulation, without a time lag, because you don't have to stop imaging when using laser light for stimulation," points out Martin Tewinkel, business manager at Olympus Life and Material Science Europa in Hamburg, Germany. "Such studies of rapid cellular responses can provide important insights into how various cellular mechanisms operate."

Another high-speed, high-resolution confocal imaging system is the Nipkow-disk-based Ultraview ERS from PerkinElmer of Boston, Massachusetts, which offers a choice of cameras for different applications: a standard interline CCD detector for the highest resolution with slower processes or bright samples that withstand higher laser intensities, or an electron-multiplying CCD detector for

AN ILLUMINATING BREAKTHROUGH

As more biologists become interested in imaging the big picture, a new microscopy technique is being developed that will enable them to look much deeper into living organisms than ever before.

Selective plane illumination microscopy (SPIM) is a timely breakthrough, according to physicist Jan Huiskens, now at the University of California, San Francisco. Huiskens co-developed the technology with colleagues when he was at the European Molecular Biology Laboratory in Heidelberg, Germany. "People are moving away from looking at single cells in a Petri dish to studying whole organs in an intact embryo," he says.

"In biology there has always been a limit — you couldn't image whole systems from living embryos," says Huiskens. With SPIM, it is possible to



Jan Huiskens: plane illumination means microscopy can now go live.

completely image living specimens 2–3 millimetres in size. The process of organogenesis — the formation of organs such as eyes and brain — can be followed, or gene and protein expression patterns tracked over days. The system can also track three-

dimensional cell cultures over time, and is fast enough for researchers to watch the heartbeat of a living fish. "The fish heart beats two to five times a second. You have to be fast enough to record that movement without blurring the image, and a high-speed camera in SPIM shooting 70 frames a second can achieve it," says Huiskens.

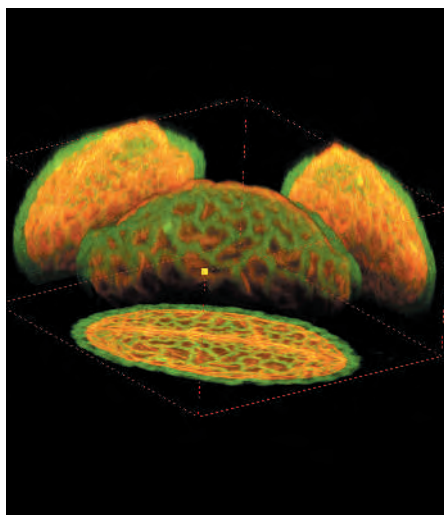
The illumination in SPIM comes from a sheet of laser light 2–8 micrometres thick, which optically sections the sample while micromotors systematically move and turn it in different directions to illuminate successive planes. The fluorescence-detection system is at right angles to the light sheet.

For developmental research, the successive layer-by-layer images are merged by an image-processing algorithm to produce three-dimensional movies of growing embryos.

SPIM has higher and more isotropic resolution compared with scanning technologies and can achieve greater penetration depth. Whereas confocals can only reach 100 to 200 micrometres deep, SPIM resolves structures inside the fish embryo at 500 micrometres and deeper when more views are combined.

Another advantage is that the embryos being studied can be kept alive in a medium-filled chamber for a few days while under observation; and the chamber is ideal for adding drugs to follow their impact on the embryo.

Although Huiskens says that it would be possible to build your own SPIM system, for those who don't feel up to tackling the job Carl Zeiss of Jena, Germany, will be commercializing this remarkable new technology within the next three years. L.M.

**3D rendering with the Olympus Fluoview.**

highly dynamic processes, very dim fluorescence, or light-sensitive samples.

For researchers who want speed but don't need the high resolution of a confocal microscope, Olympus has designed live-cell imaging systems that can be fitted to Olympus's upright BX or inverted IX series wide-field microscopes — the relatively inexpensive cell[^]M imaging station and the high-end station cell[^]R. The cell[^]R takes ten multicolour images per second at full resolution and can be used to track cell growth, metabolic transport and signal transduction in real time. "It's always a trade-off: the quality of the data on the one hand and speed on the other," explains

Christian Seel, head of information transfer management at Olympus BioSystems in Munich. The key to speed is finely synchronized illumination and camera controls. The high-intensity coloured light is switched off the specimen immediately after taking each photo, reducing photo-damage to a minimum, says Seel. Time-lapse series are stored by the imaging software and presented as movies or charts.

Developers at Carl Zeiss believe there is no need to sacrifice resolution for ultra-fast cellular dynamics with the confocal imaging system LSM 5 LIVE. "Our main motivation was to develop an instrument dedicated to fast live-cell imaging, with a much higher speed than any other system available today while maintaining a very good confocal resolution," says Richard Ankerhold, director of advanced development at Carl Zeiss in Jena, Germany. Dynamic interactions can be viewed at different scales — a group of interacting molecules, a complete cell, a developing organ, or even an entire zebrafish embryo, notes Ankerhold. LSM 5 LIVE operates even on weakly fluorescent specimens and collects up to 120 confocal images per second at a resolution of 512×512 pixels, scanning about 20 times faster than a traditional confocal system.

Developmental biologist Mary Dickinson and her group at the California Institute of Technology in Pasadena have capitalized on the instrument's fast-frame recording to image erythroblasts rushing through the heart of an 8-day-old mouse embryo, and to produce a time-lapse series of the beating heart of a

zebrafish embryo. This technology is expensive, however, and at present only affordable by large research institutes. Alternative optical approaches to imaging embryos for developmental research are selective plane illumination microscopy and optical projection tomography microscopy (see 'An illuminating breakthrough', page 775, and 'Optical tomography for embryos', below).

With calcium-sensitive dyes such as Fura-2 you can visualize spikes, waves and oscillations of calcium in living cells. An instrument dedicated to imaging intracellular ion kinetics — and with a price tag within the reach of most research labs — is the InCyt Imaging system from Intracellular Imaging of Cincinnati, Ohio, co-founded by Eric Gruenstein, director of the Center for Image Analysis at the University of Cincinnati.

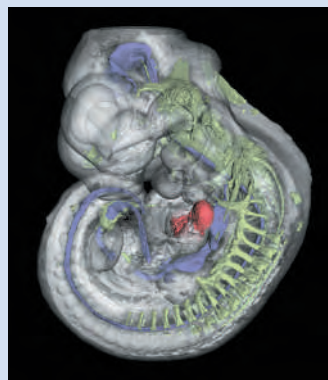
Starting at US\$30,000, it includes a fluorescence microscope, low-light-level CCD camera, filter changer and image-processing computer. "It's a very cost-effective and feature-rich solution for ion imaging and cell kinetics," says Tim Fletcher, product specialist at Image Solutions of Preston, the UK distributors for Intracellular Imaging. InCyt has been tailored to image and quantify calcium, but can measure other ions, and works with all commonly used dyes. Designed by cell biologists, the software follows the logical flow of an experiment. Images can be displayed in real time or saved and played back as an animated sequence. "The InCyt system needs very little training, and researchers pick it up quickly," says Fletcher.

OPTICAL TOMOGRAPHY FOR EMBRYOS

Developmental biologist James Sharpe became acutely aware of the drawbacks of traditional imaging methods when trying to map three-dimensional gene and protein expression patterns in the 1990s. Reconstructing images from hundreds of serial sections is not only laborious but inevitably introduces distortions. "Even in an expert lab, it takes weeks of work to reconstruct a single specimen," Sharpe points out.

To get over the problem Sharpe, who is at the MRC Human Genetics Unit in Edinburgh, UK, developed a 3D-imaging system, which he called optical projection tomography (OPT) microscopy. This approach follows the principles of the X-ray CT scanner but uses light, and so works in a very different way from confocal microscopy or SPIM, which directly image a series of optical sections. OPT is carried out on fixed specimens, and it takes 5 to

10 minutes to image an entire embryo. The specimen is rotated while the images are taken. "This new technique works in conjunction with normal microscope optics, making it cheaper than confocal and much cheaper than microscopic magnetic resonance imaging [mMRI], and therefore accessible

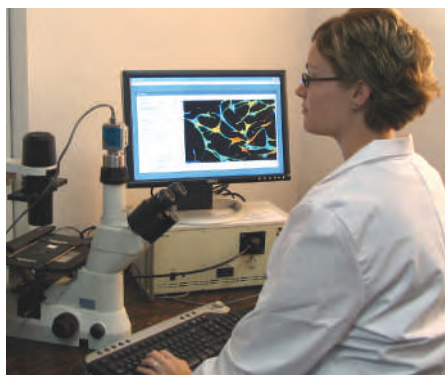
**OPT reveals the nervous system (green) and gut (blue) in a mouse embryo.**

to many more biologists", says Sharpe, who helped to found Bioptronics, the vehicle for the commercialization of OPT technology run by MRC Technology, the UK Medical Research Council's technology transfer branch. The Edinburgh-based company offers an OPT scanning service and is developing the scanners, which will be ready for sale early next year.

OPT can produce fluorescent or transmission three-dimensional images of specimens between 1 and 15 millimetres in size and computationally sections the images to reveal internal structures. Mouse, rat, chicken and zebrafish embryos fit the OPT size bracket, as does the adult fruitfly. Staining methods such as LacZ and alkaline phosphatase, which produce a non-fluorescent coloured stain, can be visualized with OPT, as well as fluorescent signals and unstained embryos.

"It fills an imaging gap by allowing analysis of specimens that are too large for confocal but too small for MRI," Sharpe says. "In addition, because it is an optical technique, OPT can extract more detail from tissues than mMRI, which cannot take advantage of fluorescent gene labelling techniques." This optical bonus can be exploited for applications outside developmental biology, including phenotyping and gene-expression analysis in adult mouse tissues.

The principle of projection tomography had not been used for optical microscopy, and there were no commercial machines. "I built the machine using parts from an old cell sorter and other equipment lying around the lab," says Sharpe, who also developed the software to turn the raw data into a 3D view of the embryo that can be rotated and sectioned to show different aspects. L.M.



Calcium signals: the InCyt imaging system.

The whole animal

For some imaging applications, getting a picture from inside the living body is the goal (see 'Optical biopsies', below). And for others, a shift from imaging *in vitro* to the whole living animal is desirable. "Many pathways interact on a systems level, that is, the immune system and the endocrine system," says Pam Contag, co-founder of imaging company Xenogen in Alameda, California. Xenogen puts together non-invasive optical imaging systems with transgenic mice and rats containing the required luciferase-tagged genes. The Xenogen IVIS 200 will image either bioluminescence or fluorescence, and its adjustable field of view makes it versatile enough to image single cells at high resolution or up to five anaesthetized mice at a time. Biophotonic imaging has proved a success with the pharmaceutical industry to test drug candidates and make more educated

predictions. "In drug discovery, the cell is used as the gold standard, but you don't treat single cells, you treat the whole person," Contag insists. "Our goal is to make animal models to be predictive for what happens in humans."

Lighttools of Encinitas, California, concentrates on whole-body imaging of animals carrying fluorescent-tagged genes. "We can zoom in and out from a couple of centimetres to 10–15 centimetres in the field of view," says John Fox, Lighttools's president. The company has recently introduced two novelties that are proving popular. One is a device for simultaneously viewing green and red fluorescent proteins (GFP and RFP) in transgenic animals, with independent controls for each fluorophore. "It allows you to turn up the RFP excitation and turn down the GFP — which is usually brighter — to obtain a more balanced image," says Fox.

The second is the Pan-A-See-Ya panoramic imaging system, which gives a 270° view of the animal in one evenly illuminated image. "It's like being able to see round the corner," says Fox. A target such as a fluorescent tumour can be viewed from two different angles at the same time with a single camera. The process is even fast enough to allow animals to be imaged without anaesthesia.

Breaking the resolution barrier

Abbe's law, postulating that optical resolution is impossible below 200 nanometres, went unchallenged for 120 years. Until recently, that is, when physicist Stefan Hell, a director of the Max Planck Institute for Biophysical

Chemistry in Göttingen, Germany, established a new law that promises greater resolution in fluorescence microscopy. The first commercial application of his new ideas is the 4Pi fluorescence confocal imaging system, created in cooperation with Leica Microsystems in Mannheim, Germany.

The Leica TCS 4Pi improves resolution of fluorescent images to 110 nanometres along the z axis. Martin Hoppe, marketing manager for Leica Microsystems, says that the system addresses a resolution gap between optical and electron microscopes. "That's what I see when I offer this system to researchers," he says. Structures down to 110 nanometres can now be resolved in living cells — a malaria parasite can now be localized precisely inside a red blood cell, for example.

For life-sciences researchers, the advantage of the TCS 4Pi over electron microscopy is that specimens can be kept alive and fluorescent stains can be used. "People don't have to redo their staining techniques," says Hoppe.

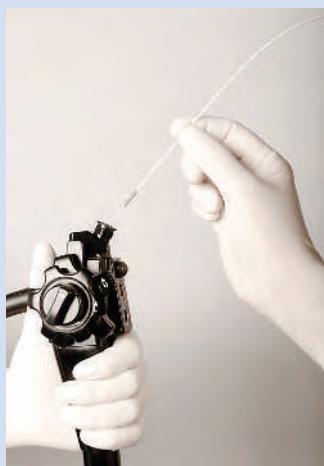
The remarkable gain in sharpness is due to the use of two opposing lenses with high numerical aperture to illuminate a single focal spot. The two wavefronts of the opposing beams interfere constructively at the focal point, which gives much higher resolution.

But at US\$1,000,000, the TCS 4Pi doesn't come cheap. "What drives up the cost is the combination of precision mechanics, interferometer optics and state-of-the-art electronics. Also, the 4Pi objectives have to be paired, this makes the manufacturing yield very low," Hoppe points out.

OPTICAL BIOPSIES

One approach to *in situ* observation in the living body is the Cell-vizio miniaturized fibre-optic confocal fluorescence microscopy system developed by Mauna Kea Technologies of Paris, France. The core technology is the equivalent of a confocal microscope on the tip of a flexible fibre optic. "It's the smallest microscope in the world: less than one-third of a millimetre in diameter," says Benjamin Abrat, general manager and co-founder of Mauna Kea Technologies. This degree of miniaturization allows *in situ* exploration of peripheral nerves, blood capillaries, internal organs such as the colon or bladder, or deep in the brain, with minimal disturbance.

Cell-vizio can image subcellular structures and at 12 frames a second is fast enough to observe cellular dynamics. Its excitation wavelength of 488 nanometres is compatible with a range of



Fibre-optic microscopes like the Cell-vizio can look inside organs.

fluorophores and with GFP and YFP transgenic animals. Used at present for small-animal research, Cell-vizio is currently being evaluated for clinical use in humans, and the company is working on European Union CE

Mark and US Food and Drug Administration approval. "The most promising applications are in cancer diagnosis," says Abrat. "The technology can enable optical biopsies — seeing at the cellular level without removing tissue," he says. Conventional excisional biopsy suffers from high false negatives due to sampling error. "With our system, you can precisely locate diseased areas," Abrat points out. The miniaturized optical catheter could simply be placed into the operating channel of an ordinary endoscope.

Optical coherence tomography (OCT) is another powerful imaging technology that can function as an 'optical biopsy'. It uses near infrared light, which reflects off the internal microstructure of biological tissues similarly to ultrasound, but gives far higher resolution, and is already in clinical use in ophthalmology for diagnosing glaucoma. With the

aim of extending OCT to other clinical applications, LightLab in Westford, Massachusetts, has designed a system that delivers infrared radiation to the target site through a single optical fibre, which could be threaded alongside a catheter. In collaboration with the University of Maryland, Baltimore, LightLab's research programme is expanding into cancer, in particular oesophageal and pulmonary cancer. The devices are being tested in animals for spinal cord imaging and the placement of deep-brain stimulation electrodes for treating Parkinson's disease (see *Nature* 436, 18–19; 2005). "The overall aim of OCT is to replace the biopsy. But we are not there yet," says Joseph Schmitt, chief technological officer at LightLab. "For the moment, the first step will be to guide the biopsy, to guide the resection, and to help make therapeutic decisions." L.M.

A world of colour

Quantum dots may be the newest kid on the block (see 'Quantum dots keep on glowing', below), but since its cloning a decade ago, the green fluorescent protein has become one of the most powerful molecular tools in the cell biology tool-box. Either by itself or as part of a fusion protein, GFP is used to visualize proteins inside living cells in a vast number of applications, from detecting gene expression to tracking cell fate in developing embryos.

The range of fluorescent proteins from jellyfish now includes red (RFP), cyan (CFP) and yellow (YFP) relatives of GFP. CFP and YFP, in particular, make a suitable contrasting pair for multicolour imaging for differential gene expression and protein localization.

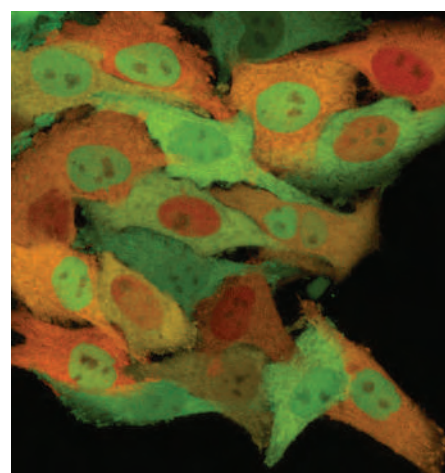
Species other than jellyfish are now being mined to expand the colour palette. The DsRed-Monomer fluorescent protein, an engineered variety of a protein from the sea anemone *Discosoma*, was recently launched by BD Biosciences Clontech of Mountain View, California, now part of Shiga-based Japanese life-sciences supply company Takara Bio.

"What is driving users' interest towards red is a better signal-to-noise ratio, because the background fluorescence from the culture medium is in the green range," explains Andrew Farmer, director of cellular and molecular biology for Clontech Business Research. The monomeric form of the new protein is an added advantage, particularly for subcellular labelling. "The DsRed-Monomer is less likely to misbehave than the tetrameric reds, which can sometimes disrupt the function of the fusion protein," says Farmer.

Stony corals have yielded a remarkable collection of new fluorescent proteins. The CoralHue range was originally isolated by Atsushi Miyawaki at the RIKEN Brain Science Institute in Saitama, Japan. Kaede (Japanese for maple leaf), the first of the family, is a brilliant green fluorescent protein that changes colour to a stable red when exposed to a short pulse of ultraviolet or violet light. Miyawaki's team have used Kaede to study hippocampal neuronal connections. Cultured neurons are labelled with green Kaede by gene transfection then, with a focused violet light pulse, a single cell body is illuminated. As the red spots spread rapidly throughout the cell's cytosol, all the nerve-cell processes, including an axon and illuminated dendrites, stand out from the green background, delineating the neuron and its multiple contact sites.

A recent addition to the CoralHue family is Kusabira orange, the first monomeric true-orange fluorescent protein. "Its greatest value is in combination with Midoriishi-Cyan for fluorescence resonance excitation transfer analysis," says Suzan Oberle, product manager for MBL International of Woburn, Massachusetts, which distributes the CoralHue range. The CoralHue pair is brighter and shows better spectral separation of donor and acceptor signals than the widely used CFP and YFP pair.

For FRAP and FLIP applications, CoralHue Dronpa green bleaches following excitation at 500 nanometres but completely regains its bright green fluorescence after minimal irradiation at 400 nanometres, without losing signal intensity. The switching can be repeated without losing brightness. Miyawaki's group



Violet light makes HeLa cells expressing Kaede fluorescent protein change from green to red.

R. ANDO & H. MIZUNO

has used this reversible protein highlighting technique to track proteins shuttling across the nuclear membrane after cell stimulation.

Another photo-switchable fluorescent protein is produced by Evrogen, a biotechnology company based in Moscow, Russia. Their cyan-to-green photo-converting protein PS-CFP2, gives a 2,000-fold increase in the green-to-cyan fluorescence ratio, making it the highest-contrast monomeric photoactivatable fluorescent protein so far.

The microscope may be 400 years old, but microscopy is refusing to show its age. High-resolution live imaging is giving researchers a whole new look at the biological world. ■

Lisa Melton is science writer in residence at the Novartis Foundation in London.

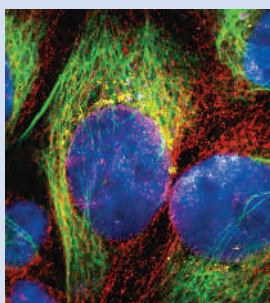
QUANTUM DOTS KEEP ON GLOWING

Further developments in imaging may well hinge on extending the range of labels. Fluorescent semiconductor nanocrystals — 'quantum dots' — are one answer. In 1998, Marcel Bruchez left the University of California in Berkeley to co-found Quantum Dot Corporation in Hayward, California. "Quantum dot nanoparticles solved a lot of the existing problems in fluorescence detection," he says. They fluoresce up to 100 times more brightly than organic dyes and are very stable. They can be fine-tuned to glow in almost any colour, each colour emitting in a narrow spectral window and thus avoiding colour overlap. And all quantum dot colours can be excited simultaneously with a single light source.

The company has just launched the Mosaic Assay System, an imaging system that uses combinations of its quantum dots

to analyse the simultaneous expression of 100 genes in high-throughput mode. A similar system for cellular and protein analysis is promised soon. Quantum Dot's largest customers are in drug discovery, but business in academia is flourishing. "The technology has captured people's imaginations. A lot of researchers have incredibly innovative ideas of what they can do with it," says Bruchez.

Tracking cancer is one example. At the Beth Israel Deaconess Medical Center in Boston, quantum dots are being used in preclinical studies in pigs to map sentinel lymph nodes. "It is possible to label the cancer and see where the



Quantum dots provide long-lasting colour.

cancer cells spread and where they stop in the lymphatic system," says George Dunbar, chief executive officer at Quantum Dot. The next step is to test the long-term safety of these materials in animals to determine whether

they could be used in people.

This question is also being addressed by quantum-dot developer Evident Technologies of Troy, New York. The latest version of its EviTag quantum dots has an outer lipid layer to help ensure biocompatibility. Another useful property of quantum dots is that they can be tailored to glow in the near infrared, allowing scientists to see right through the tissue. "The colours last longer than

organic dyes in the near infrared, and near infrared transmits through tissue without scattering," says Steven Talbot, Evident's chief marketing officer.

Evident's new indium phosphide deep red EviTags can be tuned to emit at near infrared wavelengths, so that light absorption from neighbouring tissue is avoided and reasonable signal-to-noise images generated.

Using a 4000IS scanner from Kodak of Rochester, New York, Evident has developed a label suitable for a dual-mode optical and X-ray imaging system. "You can see the bones of the mice, but also the fluorescence coming out from the quantum dots attached to the tumour cells," says Talbot. Although the question of toxicity is always on biologists' minds, Talbot is reassuring: "Several groups have been studying cytotoxicity in live cells and haven't yet seen any bad effect." L.M.

COMPANY	PRODUCTS/ACTIVITY	LOCATION	URL
Optical and fluorescence microscopy			
Amnis Corporation	ImageStream flow cytometry and cell-imaging system	Seattle, Washington	www.amnis.com
Andor Technology	CCD and intensified CCD camera systems for bioimaging	Belfast, UK	www.andor-tech.com
Applied Precision	CellWoRx for imaging living cells, SoftWoRx Explorer 5D image viewer	Issaquah, Washington	www.appliedprecision.com
Apogee Instruments	iXON electron-multiplying (EM) CCD camera and low-light imaging systems	Roseville, California	www.ccd.com
Applied Imaging	Imaging systems and image analysis for cytogenetics and pathology	San Jose, California	www.aicorp.com
Applied Scientific Instrumentation	Automated systems for microscopy and fluorescence screening	Eugene, Oregon	www.asiimaging.com
BD Biosciences Clontech	Fluorescent proteins, reagents for molecular and cell biology	Mountain View, California	www.clontech.com/clontech
Berthold Technologies	NightOWL imaging system for microscopic and macroscopic luminescent and fluorescent signals	Bad Wildbad, Germany	www.bertholdtech.com
Biomedical Photometrics	MACROscope wide-field confocal laser-scanning microscopes	Waterloo, Ontario	www.confocal.com
● Bioptonics	Optical projection tomography (OPT) scanning service	Edinburgh, UK	www.bioptronics.com
Bruker	Automated platforms for mass spectrometry, NMR, EPR, and magnetic resonance imaging	The Woodlands, Texas	www.bruker.com
Cambio	STARFISH chromosome paints, resellers of specialized biochemicals	Cambridge, UK	www.cambio.co.uk
Carl Zeiss	LSM 5 LIVE confocal laser-scanning microscopes, stereomicroscopes, upright microscopes, inverted microscopes, fluorescence correlation spectroscopy	Jena, Germany	www.zeiss.com
Chemicon	Fluorescent proteins, antibodies	Temecula, California	www.chemicon.com
Chroma Technology Corp	Optical filters and fluorescence filter sets	Brattleboro, Vermont	www.chroma.com
Cisbio international	Reagents and kits for TR-FRET (time-resolved fluorescence resonance energy transfer) assays	Bagnols-sur-Cèze, France	www.htrf-assays.com
Cytocell	Chromoprobe Multiprobe system for fluorescent <i>in situ</i> hybridization	Cambridge, UK	www.cytocell.com
DVC	Digital cameras for light microscopy, image-analysis software	Austin, Texas	www.dvco.com
Enamine	Fluorescent dyes	Kiev, Ukraine	www.enamine.net
Energy Beam Sciences	Products for light microscopy and electron microscopy	East Granby, Connecticut	www.ebsciences.com
Evident Technologies	Indium gallium phosphide MP-T2 EviTags and Biotin EviFluo quantum dot fluorescence labels	Troy, New York	www.evidenttech.com
Evrogen	Products for genomics and proteomics research, fluorescent proteins	Moscow, Russia	www.evrogen.com
Exiqon	Fluorescence-tagged LNA oligonucleotides	Vedbaek, Denmark	www.exiqon.com
Genisphere	3DNA FISH probes	Hatfield, Pennsylvania	www.genisphere.com
● Hamamatsu Photonic Systems	Digital imaging systems and cameras	Hamamatsu City, Japan	www.hamamatsu.com
Image Solutions	Microscopes and imaging systems	Preston, UK	www.imsol.co.uk
Intracellular Imaging	Digital fluorescence and photometry systems	Cincinnati, Ohio	www.intracellular.com
Intelligent Imaging Innovations	Digital microscopy workstations and imaging software	Denver, Colorado	www.intelligent-imaging.com
Kodak	Image Station <i>in vivo</i> F and FX multi-mode imaging systems, image-analysis software	Rochester, New York	www.kodak.com
LaVision Biotec	TriMScope laser-scanning microscope, fluorescence-lifetime microscopy system	Bielefeld, Germany	www.lavisionbiotec.de
● Leica Microsystems	TCS 4Pi fluorescence confocal imaging system, DM digital microscope range, microscope systems	Wetzlar, Germany	www.leica-microsystems.com
LightLab	Optical coherence tomography (OCT) development	Westford, Massachusetts	www.lightlabimaging.com
Lighttools	Whole mouse imaging, Pan-A-See-Ya panoramic imaging system	Encinitas, California	www.lighttools.com
Martek Biosciences	Fluorescent tags based on algal proteins	Columbia, Maryland	www.martekbio.com
Mauna Kea Technologies	Miniaturized fibre-optic fluorescence confocal microscopy system for <i>in vivo</i> intra-tissue imaging	Paris, France	www.maunaakeatech.com
MBL International	CoralHue range of fluorescent proteins	Woburn, Massachusetts	www.mblintl.com
Micro Luminetics	Cryocam cooled CCD cameras for low-light imaging	Los Angeles, California	www.cryocam.com
Micro Video Instruments	Microscopes and accessories	Avon, Massachusetts	www.mvi-inc.com
Nanoptek	Digital photon-tunnelling microscope with near-field optics	Concord, Massachusetts	www.nanoptek.com
● Nikon Instruments	Digital Eclipse C1 Spectral Imaging confocal system, microscopes	Melville, New York	www.nikon.com
Novagen	Reagents and kits for molecular biology, SensiLight fluorescent reagents	Madison, Wisconsin	www.novagen.com
● Olympus Life and Material Science Europa	Digital microscopes, wide-field microscopes, Fluoview confocal laser-scanning microscopes, cell range of imaging stations	Hamburg, Germany	www.olympus.com
Omega Optical	Optical filters and fluorescence filter sets, microscopes	Brattleboro, Vermont	www.omegafilters.com
Open Biosystems	STAR [®] FISH chromosome paints	Huntsville, Alabama	www.openbiosystems.com
Optical Imaging	Brain and retinal imaging systems	Rehovot, Israel	www.opt-imaging.com
Optronics	Digital microscope cameras, image-analysis software	Goleta, California	www.optronics.com
PCO	Digital video camera systems	Kelheim, Germany	www.pco.de
PerkinElmer Life Sciences	UltraView confocal laser-scanning microscope system	Boston, Massachusetts	las.perkinelmer.com
Photon Technology International	Imagemaster fluorescence and general imaging systems	Birmingham, New Jersey	www.pti-nj.com
Photometrics	CCD and EMCCD cameras	Tucson, Arizona	www.photomet.com
Polaroid	Digital imaging systems for microscopes	Waltham, Massachusetts	www.polaroid.com
Princeton Instruments	Equipment for imaging and spectroscopy	Trenton, New Jersey	www.princetoninstruments.com
Qbiogene	Fluorescent proteins and fusion vectors, products for FISH	Irvine, California	www.qbiogene.com
QImaging	High-performance digital cameras and RGB filters for microscopy	Burnaby, British Columbia	www.qimaging.com
● Quantum Dot Corporation	Qdots quantum dot fluorescent tags, Mosaic gene expression assay system	Hayward, California	www.qdots.com
Richardson Technologies	RTM 3.0 microscope system, high-resolution field microscope	Bolton, Ontario	www.richardson-tech.com
Shimadzu North America	AIM 8800 automated Fourier transform infrared microscope	Columbia, Maryland	www.ssi.shimadzu.com
Thermo Electron	Infrared and Raman microscopes, laser sources, image-analysis software	Waltham, Massachusetts	www.thermo.com
Tritech Research	Dissecting stereomicroscope system, microscope accessories	Los Angeles, California	www.tritechresearch.com
Unitron	Optical microscopes	Bohemia, New York	www.unitronusa.com
Veeco	Scanning-probe microscopes	Woodbury, New York	www.veeco.com
WITec	Optical and scanning-probe microscopes	Ulm, Germany	www.witec.de
Varian	Spectrophotometers, infrared microscopes and imaging systems	Palo Alto, California	www.varianinc.com
VayTek	Digital imaging systems, cameras, imaging software	Fairfield, Iowa	www.vaytek.com
Vision Engineering	Microscopes, microscope accessories	Woking, UK	www.visioneng.com

COMPANY	PRODUCTS/ACTIVITY	LOCATION	URL
Vysis	Chromosome paints for FISH	Downer's Grove, Illinois	www.vysis.com
Xenogen	Luciferase-based <i>in vivo</i> imaging for live animals	Alameda, California	www.xenogen.com
Laser scanners and imaging systems			
Affymetrix	GeneChip microarray systems and scanners	Santa Clara, California	www.affymetrix.com
Agilent	Catalogue and custom DNA microarrays, scanner and software	Palo Alto, California	www.agilent.com
Alpha Innotech	Alpha Array microarray reader	San Leandro, California	www.alphainnotech.com
Applied Biosystems	DNA sequencers, protein sequencers, mass spectrometers, chromatography systems	Foster City, California	home.appliedbiosystems.com
Beckman Coulter	Automated tools for molecular biology, biochemistry, genomics and proteomics	Fullerton, California	www.beckmancoulter.com
● Bio-Rad	Imaging systems for molecular biology	Hercules, California	www.bio-rad.com
BMG Labtech	Scanners, microarray readers and handling systems	Offenburg, Germany	www.bmg-labtechnologies.com
Cellomics	ArrayScan HCS and KineticScan HCS systems for automated cell-based high-content screening	Pittsburgh, Pennsylvania	www.cellomics.com
CyBio	Pipetting, liquid handling, incubation, and imaging systems, CyBio Nanoscan plater readers	Jena, Germany	www.cybio-ag.com
Cyntellect	High-throughput LEAP laser-scanning and image-analysis system for cellular analysis	San Diego, California	www.cyntellect.com
Fujifilm	Imaging systems for protein electrophoresis and array analysis	Stamford, Connecticut	www.fujimed.com
● GE Healthcare	IN Cell confocal imaging system for cell screening, scanners and analysis software	Chalfont St Giles, UK	www.amersham.co.uk
Genetix	Q-bot scanners	New Milton, UK	www.genetix.co.uk
LI-COR Biosciences	Infrared imaging system for western blot and in-cell western blot assay analysis	Lincoln, Nebraska	www.licor.com
Metrigenix	4D Assay scanner for DNA microarrays	Gaithersburg, Maryland	www.metrigenix.com
MiraBio	CRBIO lie scanner, image-analysis software	Alameda, California	www.mirabio.com
Molecular Devices	GenePix 4000B microarray scanner and analysis software, ImageExpress for live-cell assays	Foster City, California	www.moleculardevices.com
UVP	Bioimaging systems for gels, microarrays and microplates	Upland, California	www.uvp.com
Image-analysis software			
BioImage	CellMine image-analysis software for high-content screening	San Mateo, California	www.bioimage.com
Bitplane	Image-processing software	Zürich, Switzerland	www.bitplane.com
Clemex	Image-analysis software	Longueuil, Quebec	www.clemex.com
Definiens	Cellenger image-analysis software	Munich, Germany	www.definiens.com
Improvision	3D imaging software	Coventry, UK	www.improvision.com
Media Cybernetics	ImagePro imaging software	Silver Spring, Maryland	www.mediacy.com
Mirero	Image-analysis software	Gyeonggi-do, South Korea	www.gaia-zone.com
Scanalytics	IPLab scientific imaging software	Rockville, Maryland	www.scanalytics.com
Wavemetrics	IGOR image-processing and analysis software	Portland, Oregon	www.wavemetrics.com
General			
Active Motif	Kits and reagents for nucleic acid isolation, cell fractionation and gene silencing	Carlsbad, California	www.activemotif.com
Aetos	Biopolymer for preserving and transporting live specimens	Ophelia, Alabama	www.aetos.com
Arcturus	Laser-capture dissection, microgenomics, RNA amplification	Mountain View, California	www.arctur.com
Biognostik	Antisense oligonucleotide synthesis kits and services	Göttingen, Germany	www.biognostik.com
BIOMOL	Biochemicals for life-science research	Plymouth Meeting, Pennsylvania	www.biomol.com
Biosearch Technologies	Fluorophores and quenchers, Black Hole Quencher	Novato, California	www.biosearchtech.com
Cambrex	Products for molecular and cell biology research	Walkersville, Maryland	www.cambrex.com
Cayman Chemical	Antibodies, proteins and small biological molecules	Ann Arbor, Michigan	www.caymanchem.com
EMD Biosciences	Novabiochem, Calbiochem, Novagen and Oncogene Research products	San Diego, California	splash.emdbiosciences.com
Enzo Life Sciences	Nonradioactive labelling and detection reagents for molecular biology	Farmingdale, New York	www.enzolifesciences.com
HORIBA Jobin Yvon	Fluorescence spectrophotometers	Kyoto, Japan	www.jobinyvon.co.uk
Eppendorf	Laboratory instrumentation and consumables for molecular and cell biology	Hamburg, Germany	www.eppendorf.com
● Genscript	Kits and services for molecular biology, oligonucleotide probes	Piscataway, New Jersey	www.genscript.com
Hamilton Company	Automation for the life-sciences	Reno, Nevada	www.hamiltoncomp.com
IBA	Oligonucleotide probes	Göttingen, Germany	www.ibo-go.com
KPL	Nucleic-acid labelling and detection kits	Gaithersburg, Maryland	www.kpl.com
Invitrogen	Kits and reagents for genomics, proteomics, molecular and cell biology	Carlsbad, California	www.invitrogen.com
Molecular Beacons	Molecular beacon probes for <i>in situ</i> RNA detection	Newark, New Jersey	www.molecular-beacons.org
Pierce Chemical/ Pierce Biotechnology	Products for molecular biology, cell biology, immunology and chromatography	Rockford, Illinois	www.piercenet.com
Promega	Vectors, reagents and kits for genomics, proteomics and cell biology	Madison, Wisconsin	www.promega.com
Qiagen	Automated workstations, kits and reagents for genomics, proteomics, molecular and cell biology	Valencia, California	www.qiagen.com
● R&D Systems	Antibodies, fluorescence-based cytokine detection kits	Minneapolis, Minnesota	www.RnDSystems.com
Roche Applied Science	Instrumentation and reagents for PCR	Lewes, UK	www.roche-applied-science.com
● Sigma-Aldrich	Reagents and biochemicals for life sciences research	St Louis, Missouri	sigma-aldrich.com
● Stratagene	Kits and reagents for genomics and molecular biology	La Jolla, California	www.stratagene.com
● Takara Bio	Reagents, kits and custom services for molecular biology, genomics, proteomics research	Shiga, Japan	www.takara-bio.com
Tecan	Laboratory automation for the life sciences	Männedorf, Switzerland	www.tecan.com
Upstate	Kits, reagents and services for cell signalling, neuroscience and cell biology	Charlottesville, Virginia	www.upstate.com
USB	Reagents for molecular and cell biology	Cleveland, Ohio	www.usbweb.com

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

Stemming the tide

Arnold Schwarzenegger, movie star and governor of California, helped bring stem-cell funding to the state with the much-debated Proposition 71. But it is not clear whether the \$3-billion initiative will stop stem-cell scientists leaving the United States. A panel discussing the business and financing of stem-cell research last week at the University of California, San Francisco, was split about the initiative's ability to slow the stem-cell brain drain.

The panellists, convened by BayBIO, a regional biotechnology organization and co-sponsored by *Naturejobs*, agreed that Proposition 71 was intended to keep the best stem-cell scientists in California. But they weren't sure how effective it would be.

Rik Derynck, a tissue biologist at the university, says that money from the initiative hasn't yet appeared in any research department coffers. Although the funding would be helpful, it's not vital for retaining scientists interested in stem-cell research. Much basic work doesn't rely on human embryonic stem cells, the kind restricted under US government funding rules. And private foundations and businesses will fund research they find relevant. Stacy

Taylor, a lawyer at Foley & Lardner's San Diego office, also points out that the United States is in an enviable position as regards intellectual property, at least for the early stages of applied research. Although the country is one of several with stem-cell policies that she calls "restrictive", it has the bulk of stem-cell patents at present.

But from a business perspective, this may not be enough. Antoun Nabhan, who works for a venture-capital firm in San Diego, predicts that countries with more lenient policies, such as Britain, Korea and Singapore, will gain an advantage.

Jobs in the stem-cell industry will follow the first success. Given the political climate, there is no guarantee that this success will come in the United States — even with more state-led initiatives like California's.



Paul Smaglik, *Naturejobs* editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London
 The Macmillan Building, 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs Sales Director:
 Nevin Bayoumi (4978)
European Sales Manager:
 Andy Douglas (4975)

Natureevents: Sille Opstrup (4994)
UK/RoW/Ireland/Italy:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
France/Switzerland/Belgium:
 Amelie Pequignot (4974)
Germany/Austria/The Netherlands:
 Reya Silao (4970)

Advertising Production Manager:
 Billie Franklin
 To send materials use London
 address above.
 Tel: +44 (0) 20 7843 4814

Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
Germany/Austria/
The Netherlands:
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com

US Head Office, New York
 345 Park Avenue South,
 10th Floor, New York,

NY 10010-1707
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

Capital attractions

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

A research career was just a dream for Mateusz Kolanczyk and Tomasz Zemojtel when the two kindergarten friends found themselves together at high school, and later university, in Gdansk, Poland, in the 1990s. Kolanczyk's interest in biology had been sparked by a popular article he once read about the Human Genome Project, but he also remembers how his friend had to help him out in chemistry, where he had miserable marks.

Miracles do happen. After many twists and turns, the two met again last year at the Max Planck Institute (MPI) for Molecular Genetics in Berlin, one of Europe's leading research institutes in the field. Now postdocs — in molecular biology and bioinformatics, respectively — they soon discovered a common interest in a mitochondrial gene relevant for bone biology. Thus, some 15 years after they first started to enthuse about the wonders of life, they are set to embark on a research project of their own.

"We always wanted to do something together," says Zemojtel. "Now it seems I can deliver the hypothesis, and Mateusz will test it in the wet lab. It's just fantastic."

It could have happened anywhere, but Germany's capital is perhaps predestined for such a saga. An island for almost 30 years during the cold war, Berlin has reinvented itself in the past 15 years as the melting pot of people and cultures that it used to be in the early twentieth century. Today, the city is an unlikely mix of the nostalgic and the futuristic, the trivial and the sublime, a matchless memorial of history and a bold promise of tomorrow — and a place changing so fast that it is hard to keep pace.

Since the wall fell in November 1989, Berlin has also become a hot spot for science, particularly the life sciences. The transformation of the former communist

Despite funding uncertainties, Berlin's facilities, charisma and cosmopolitan atmosphere continue to draw researchers from across Europe, says Quirin Schiermeier.



Academy of Sciences and Humboldt University, and the creation of a number of research institutes and technology parks, was achieved despite the notorious budget crisis that has gripped the city since reunification, when cold-war subsidies were stopped.

Packing them in

The Charité, the renowned medical school and hospital of Humboldt University, is at the heart of the debate about how much science the city can afford. The centre, whose four campuses are spread across the city, has to save almost a quarter of a billion euros (US\$300 million) by 2010, more than 20% of its budget. Its centrepiece, a high-rise block built in 1978, was once the symbol of medicine in East Germany. The building is urgently awaiting renovation, but some people in Berlin would rather tear down the legacy of communism.

Learning to get by is not easy for scientists and science managers in Berlin, where some complicated east-west sensitivities still prevail. On the credit side, they can rely on an inexhaustible influx of scientific talent into the city — in many regards the guarantor of Berlin's scientific success story.

"Berlin's charisma is a major pulling power," says Detlev Ganten, chairman of the Charité.

Clemens Schmitt says the Berlin effect has helped him build up a 20-strong team quickly. In 2001, the molecular oncologist returned from a postdoc at Cold Spring Harbor Laboratory, New York. He rejected an offer from the Memorial Sloan-Kettering Cancer Center in New York and decided to work in his native Germany, where he now holds a joint group-leader

Reunited: Mateusz Kolanczyk, left, and Tomasz Zemojtel.

U. SCHMID/CORBIS

Q. SCHIERMEIER

position at the Charité and the Max Delbrück Centre of Molecular Medicine (MDC).

He doesn't regret the decision.

"The width of institutes here, and the strong network of disciplines, makes it easy to find new partners," he says. "Newcomers to Berlin really benefit from the great potential for recruitment here."

A good part of the region's science base is located on Berlin's periphery: the Free University, founded in 1948, is spread all over Dahlem, a neat residential area in the southwest of the city. Potsdam, the nearby capital of the surrounding state of Brandenburg, hosts its own university and a number of high-profile astronomy, climate and geosciences institutes in the historic Albert Einstein science park.

At the eastern end of Berlin are two large science parks: Adlershof, which hosts most of Humboldt University's science departments, and Berlin-Buch, a life-science, health and biotechnology campus renovated and expanded after reunification.

Close to home

Maciej Lalowski, a Polish neurobiologist at the MDC, says he likes the "humane, European face" of the German science system, which he finds less cut-throat, but not necessarily less efficient, than systems in places such as Boston or New York.

"I was smiling when I came here," he says. "I was feeling at home. But then it was also kind of weird to see all the Soviet-style low-budget buildings around the area. For my Finnish wife it was really a shock."

Like Lalowski, many scientists from central and eastern Europe have chosen Berlin partly because it is close to their home countries.

The Polish border is just 80 kilometres to the east, and Prague and Budapest are within reach. "It's a



Q. SCHIERMEIER



Q. SCHIERMEIER

K. STRAITON/CORBIS

Members of Clemens Schmitt's team: "Newcomers benefit from the great potential for recruitment here."

weekend option to visit my family in Poznan," says Kolanczyk. "Berlin also has nice low rents and there are many cheap airlines. I see my Polish friends much more often now than I would at home. They often stay overnight at my place before travelling abroad."

"I go home to St Petersburg every three to four weeks," says Natalia Alenina, a Russian stem-cell researcher at the MDC. For young scientists like her, who often leave their country without speaking a word of German, and little English, it is easier to assimilate in Berlin than it would be in, say, a small university town. "I like the culture, the nightlife, the people. It's this cocktail that makes Berlin so exciting," she says.

Then, of course, there is the multitude of scientific opportunities, the Max Planck research schools, the PhD programmes, and the strong research infrastructure. Among other highlights, Berlin hosts Europe's largest animal facility at the MDC; the BESSY II synchrotron radiation facility in Adlershof; the Research Centre for Mathematics for Key Technologies; and the German Resource Center for Genome Science, a not-for-profit service centre for genomics and proteomics research.

Missed chances

But Martin Vingron, scientific director of the MPI for Molecular Genetics, believes that Berlin and Germany could have done better. "I know many Russian scientists in the United States who would prefer to work in Germany," he says. "We missed the opportunity, whereas US universities brought out the red carpet for them. We just don't have the manners needed to compete at that level."

Some newcomers do find the *Berliner Schnauze*, the notorious Berlin bluntness, offputting. "You wouldn't really expect to be yelled at by shop assistants, particularly if you have just arrived from Cambridge," says Michael Lappe, a biochemist and junior research group leader at the MPI.

But most foreign scientists, such as Kolanczyk and Zemojtel, see the friendly side of Berlin. "It's simply great being here and having met each other again," they say. "We don't regret a single moment."

Quirin Schiermeier is *Nature's* German correspondent.

LEADING INDICATORS: BERLIN

Biotechnology

156 companies in the Berlin-Brandenburg area.

Patents

13% of all German science applications.

R&D workforce

About 31,000 people — 6.6% of Germany's R&D.

Staffing

There are 200 working groups with about 5,000 members, the highest research density in Germany.

Public funding

€1.8 billion (US\$2.2 billion).

Universities

Berlin-Brandenburg has four universities with about 130,000 students, and 250 other research institutes.

Biotech parks

Seven.

Housing

Monthly rent on a three-room, 70-square-metre apartment: about €260–400, or €3.70–5.80/m².

Public transport

Day ticket for bus and metro: €6.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Visionary: Zemojtel and Kolanczyk enjoy working in the rejuvenated city, epitomized by the new dome (above) of the Reichstag (opposite).

Feeling rejected

World beater.

Alastair Reynolds

Report on the paper "Analysis of the gravitational signals from a newly discovered Kardashev II civilization in the Sombrero Galaxy: Part 1 by Whimbrel *et al.*", submitted to the *Journal of Xenoastronomical Studies*.

The authors present an analysis of gravitational signals of intelligent origin arising in the Sombrero Galaxy, detected in publicly available archival data from the System-Wide Imaging Network for Exo-astronomy (SWINE). The transmitting culture, which has not been the subject of an earlier paper, is shown to be a type II civilization on the Kardashev scheme, by which it is understood that they have the means to tap the entire energy output of their star. This classification is made partly on the basis of the strength of the SWINE signal (which in itself implies a basic competence in stellar husbandry) and partly on the basis of the cultural information embedded in the data themselves. This assessment is probably correct, but given the likelihood that both type I and type III civilizations may occasionally emulate type II civilizations for their own purposes (see, for instance, Chukar, Francolin and Dickcissel, 2051), a word of caution might well have been in order.

The species is shown to have originated on a rocky terrestrial planet about the size of Mars, and to have followed an evolutionary pathway that is well approximated by the uppermost track on the three-parameter model of Bataleur and Becard (2049). In their unmodified form, adult members of this species are 3-metre tall hexapodal oxygen-breathers with a DNA-based reproductive system. The species has a well-developed central nervous system with marked hemispheric asymmetry.

The authors apply standard analysis tools and methods to extract cultural information from the intercepted signal. Given the absence of anything startlingly new in their approach, the amount of space that the authors spend discussing this process is puzzling. It might have been better simply to reference one of the many review papers on the matter, such as the recent and comprehensive overview of analysis methods given in [omitted].

The authors then move on to the main

part of their paper: a lengthy discussion of the information content of the decoded message. They summarize the nature of the transmitting civilization, the physiology and evolutionary background of the inhabitants, their technology and culture. Although broadly satisfactory in its details, this section would benefit from shortening. As an example, the authors dwell on the construction methods used in the Dyson sphere that the aliens have erected around their star, despite the fact that broadly similar planet-dismantling, reforging and gravity-control methods have been used by at least 138 other Kardashev II cultures (see, for instance,



Takahe and Smew, 2045). In the very first sentence of subsection 3.2, the authors state that there is "nothing particularly novel about the construction methods", before nevertheless embarking on a blow-by-blow account of those selfsame methods. I agree with the first sentence.

They conclude this section by presenting, in excerpted form, several images and texts deemed to be of high significance within the culture. These include 18 'stanzas' of a much longer epic 'poem' written in commemoration of the collapse of part of the polar region of their Dyson sphere about 1.2 million years ago, an accident that resulted in the deaths of $\sim 5.6 \times 10^{12}$ sentient beings. Although undoubtedly touching, it is not clear that a great deal is gained from the inclusion of this somewhat taxing material.

The authors conclude their paper by moving on to a wider discussion of the significance of their newly found civilization against the known sample of other intelligent alien species. Here the authors place (in my view) undue emphasis on the position of their civilization in the 'cultural H-R diagram' (Wonga and Grebe, 2044),

in which the total information capacity of a transmitting culture, measured in bits, is plotted against the light-crossing time in light-seconds of their total colonized space. On the basis of Figure 8, the authors claim that their culture lies significantly to the right of the 'asymptotic singularity branch', which on the face of it would suggest that the culture had avoided a singularity despite occupying a total volume only 72,000 light-seconds in diameter. If true, this would extend the total number of known collapse-resistant cultures to eight.

The evidence, however, is very much less compelling than the authors claim. Close examination of their statistical sample shows it to be derived from the Third Gonolek catalogue, which is now known to be afflicted by serious sampling errors. Visual inspection would suggest that a more reliable sample — such as that of [omitted] — would either bring their civilization into line with the singularity branch, or reveal it as no more than a mild outlier.

In short, although the new civilization undoubtedly provides a useful new datum point, I remain unconvinced that it merits an entire paper, and certainly not the multi-paper saga that the authors clearly have in mind. Thanks to the torrent of data supplied by SWINE, and the planned Obscenely Large Gravity Array (OLGA), we are fast moving into the era of statistical xenoastronomy — one in which the study of individual extraterrestrial civilizations has much less to offer than a global, survey-based approach. The authors might therefore be advised to wait until their archival enquiries have turned up several dozen such cultures, and then gather these results into a single paper. Otherwise, I fear, they may be open to accusations of [omitted] the [omitted].

Other matters:

Fig. 6 was incorrectly labelled (see Fig. 5).

This text has been swept by Semantic Anonymity Preserver Version 5.1 — certain stylistic or cultural markers may have been altered or removed. Complete legibility is not guaranteed. ■

Alastair Reynolds lives in the Netherlands. He worked at the European Space Agency for 12 years on a variety of astronomical projects before turning to full-time writing. His latest book is *Century Rain* (Orion, 2004).

JACEY